OPREDELITEV KAKOVOSTI PODATKOV IN NJENO ZAGOTAVLJANJE V RELACIJSKEM PODATKOVNEM MODELU POSLOVNO INFORMACIJSKEGA SISTEMA

UROŠ GODNOV

Društvo za akademske in aplikativne raziskave

Koper, 2012

Izdalo in založilo: Društvo za akademske in aplikativne raziskave Koper

Recenzenta: dr. Aleš Groznik

dr. Srečko Natek

Elektronska izdaja.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

659.23:004.65

GODNOV, Uroš
Opredelitev kakovosti podatkov in njeno zagotavljanje v relacijskem podatkovnem modelu poslovno informacijskega sistema [Elektronski vir] / Uroš Godnov. - Koper : Društvo za akademske in aplikativne raziskave, 2012

Način dostopa (URL): http://www.bizis.si/

ISBN 978-961-6709-13-2 (pdf)

263877120

OPREDELITEV KAKOVOSTI PODATKOV IN NJENO ZAGOTAVLJANJE V RELACIJSKEM PODATKOVNEM MODELU POSLOVNO INFORMACIJSKEGA SISTEMA

UROŠ GODNOV

Društvo za akademske in aplikativne raziskave

Koper, 2012

PREDGOVOR

Kot projektant in programer že več kot sedem let delam na področju razvoja in uvedbe poslovnih informacijskih sistemov (v nadaljevanju PIS) v združbe. Največ pozornosti namenjam razvoju relacijskih podatkovnih zbirk v smislu njihove fizične izvedbe ter izdelavi skladišč podatkov, nadgrajenih s tehnologijo OLAP (angl. *On Line Analytical Processing*). Pri svojem delu sem opazil, da se kljub velikemu trudu pri načrtovanju, programiranju, preizkušanju in izobraževanju uporabnikov pojavljajo napake v PIS. Te zahtevajo poseg »od zadaj«, torej mimo uporabniškega vmesnika, neposredno v zbirko podatkov. Takšnih posegov ni bilo malo, še posebno so bili pogosti ob novih različicah uporabniških rešitev, ki so prinesle večje spremembe. Čeprav nam je veliko napak uspelo oziroma jih še uspevamo reševati s pomočjo uporabniškega vmesnika, ostajajo določeni problemi oziroma napake, ki zahtevajo poseg neposredno v zbirko podatkov.

Pri delu pa je bilo mogoče opaziti še eno značilnost končnih uporabnikov. Združbe, ki so imele organizirano lastno službo informatike, so mnogokrat same, brez vednosti ponudnika PIS, posegala v zbirko podatkov. To je bilo še posebno izrazito v združbah, ki so dlje časa vztrajale pri istem ponudniku in so imele znanje ter čas proučiti zgradbo informacijskega sistema, ki so ga vsakodnevno uporabljale pri svojem delu. Mnogokrat so končni uporabniki prihajali v službo za informatiko ter prosili za neposreden poseg v zbirko podatkov (na primer uporabnik se je zmotil pri temeljnici in namesto storniranja je prosil sodelavca informatika, da je temeljnico popravil neposredno v zbirki podatkov). Nekatere združbe so storile še korak dlje in so na zbirko podatkov »priključile« svoje uporabniške rešitve.

Poleg načrtovanja in razvoja sistemov sprotne obdelave podatkov sem načrtoval in razvijal tudi sisteme za spoznavanje in predvidevanje poslovanja združb s pomočjo računalniških orodij za poslovno obveščanje (v nadaljevanju poslovno obveščanje). Ti so podatke v pretežni meri črpali iz relacijskih zbirk podatkov. Kljub dobri pripravi in znanju ni noben projekt, ki sem ga vodil, uspel v načrtovanem časovnem roku. Vzrok – NEKAKOVOSTNI PODATKI. Največ časa se je porabilo prav pri ukvarjanju z nekakovostnimi podatki, ki so oteževali izdelavo sistemov za poslovno obveščanje. Mnogokrat združbe sploh niso imele védenja o stanju svojih podatkov in je projekt izdelave analitičnih rešitev pomenil neke vrste streznitev. Največ napak je bilo, po mojem mnenju, posledica slabe organizacije poslovnih procesov in pomanjkanja deklarativnih omejitev (na primer kdo je zadolžen za vnos določenih podatkov o entitetah in kateri podatki o entitetah se morajo vnašati), ki so se pokazale v podvojenih vnosih entitet, neažurnosti podatkov o entitetah, v manjkajočih podatkih in še bi lahko našteval. Ogromno teh napak se ne

bi pojavilo, če bi PIS imel kakovosten fizičen podatkovni model. in prav to me je navdušilo za knjigo – torej proučiti, kako fizična izvedba relacijskega podatkovnega modela vpliva na kakovost podatkov.

Knjiga pred vami, ki nosi naslov »Opredelitev kakovosti podatkov in njeno zagotavljanje v relacijskem podatkovnem modelu poslovno informacijskega sistema« torej govori o vplivu glavnih in tujih ključev, deklarativnih omejitev, normalizacije ter podatkovnih tipov na natančnost, doslednost in popolnost podatkov v poslovnih informacijskih sistemih.

Kakovost podatkov je v novejšem času izredno aktualno področje, ki ima svoje korenine v 80. letih prejšnjega stoletja, aktualnost pa je mogoče pripisati razširjenosti in pomembnosti poslovnega obveščanja v združbah. V literaturi obstajajo tri pomembne opredelitve oziroma pojmi, povezani z opredelitvijo kakovosti podatkov. Najbolj razširjena je statična opredelitev, ki predpostavlja sodila oziroma razsežnosti kakovosti podatkov, združene v štiri skupine. Naknadno se je pojavil še vidik namena uporabe ter dinamične opredelitve. Namen uporabe je v knjigi predstavljen kot osrednji vidik, ki ga mora posamezna združba upoštevati, in predpostavlja, da se morajo podatki ocenjevati z vidika namena uporabe. Dinamični vidik pa ocenjuje kakovost podatkov v različnih korakih procesa ravnanja s podatki. V knjigi so vsi trije vidiki združeni v celovito opredelitev.

Relacijski podatkovni model je trenutno najbolj razširjen in verjetno bo tako še nekaj časa, saj temelji na teoriji množic, torej znanosti. In stvari, ki temeljijo na znanosti, imajo po navadi daljši rok trajanja. Utemeljitelj relacijskega podatkovnega modela je Edgar Codd, ki je postavil temeljna pravila in jih pozneje tudi dopolnjeval. Relacijski model nikjer ne zapoveduje načina fizične izvedbe, vendar so vsi najpomembnejši SUPB-i sledili podobnim smernicam. Zato v knjigi med pomembne fizične lastnosti uvrščamo tudi podatkovne tipe in deklarativne omejitve.

Za proučevanje vpliva fizičnih lastnosti relacijskega podatkovnega modela na kakovost podatkov so bile izvedene dve dejavnosti, in sicer modeliranje s SUBP MS SQL 2005 ter anketiranje slovenskih združb. Z modeliranjem z MS SQL 2005 sta bila prikazana dva razmeroma preprosta scenarija, ki pa sta zajela vse proučevane lastnosti relacijskega podatkovnega modela ter najpomembnejše razsežnosti kakovosti podatkov. Proučevanje se je nadaljevalo in hkrati končalo z izvedbo anketiranja, ki je zajelo majhne, srednje velike in večje združbe vseh statističnih regij ter panog dejavnosti. Za raziskavo je bilo dobljenih 74 pravilno izpolnjenih anket, kar je bilo zadostno število za ustrezno statistično analizo, ki je dodatno potrdila rezultate modeliranja. Statistična raziskava je pokazala, da sta pravilna uporaba podatkovnih tipov in deklarativnih omejitev ter spoštovanje normalizacije najpomembnejša pri zagotavljanju natančnih, doslednih in popolnih podatkov. Raziskava je pokazala tudi na dejstvo, da večine dejavnikov, ki vplivajo na razsežnosti kakovosti podatkov, ni mogoče iskati v relacijskem

podatkovnem modelu. Ta ugotovitev je v skladu z raziskavo TDWI, ki je ugotovila, da največ napak v kakovosti podatkov izhaja iz napačnega vnosa podatkov zaposlenih.

Slovenske anketirane združbe so glede kakovosti podatkov podobne združbam iz ZDA. Deleža združb s težavami v kakovosti podatkov sta si podobna, prav tako deleža dojemanja podatkov kot pomembnega premoženja združb. Tudi pri zaznavi vzrokov za nekakovost lahko slovenske anketirane združbe postavimo ob bok združbam iz ZDA.

Čeprav obravnavana tematika ni povsem tuja oziroma nova, je nov način, kako je bila proučevana povezava med lastnostmi relacijskega podatkovnega modela in razsežnostmi kakovosti podatkov. Uporabe izkustvene metode pri proučevanju povezav namreč ni mogoče zaslediti v nobeni raziskavi.

Raziskava je torej potrdila vpliv proučevanih lastnosti relacijskega podatkovnega modela na kakovost podatkov, kar je opozorilo združbam, da morajo pri zagotavljanju kakovosti podatkov posvetiti pozornost tudi relacijskemu podatkovnemu modelu poslovnih informacijskih sistemov. Vendar je to le del celotnega procesa ravnanja s kakovostjo podatkov, ki bo moral postati del poslovne strategije vsake združbe, neodvisno od njene velikosti in panoge dejavnosti.

Pričujoča knjiga vam bo torej dala vpogled v svet kakovosti podatkov, hkrati pa vam bo utrdila prepričanje, da se vse skupaj začne že pri modeliranju zbirke podatkov poslovno informacijskega sistema.

Avtor

Kazalo vsebine:

PREDGOVOR III

1 UVOD 1

1.1 HIPOTEZE RAZISKAVE 7
1.2 SESTAVA KNJIGE 9
1.3 KAJ JE ŽE DOGNANO IN KAJ NI? 10

2 PODATKI, PROBLEMI IN ODLOČITVE 13

2.1 ZNANJE IN MODELI 13
2.2 ODLOČITVENE RAZMERE IN ODLOČITVE 14
2.2.1 Odločitveni okvir 14
2.2.2 Urejenost odločitvenega procesa 16
2.2.3 Vrsta odločanja 17
2.2.4 Odločitev in tveganje 18

3 ZBIRKE PODATKOV KOT TEMELJ SHRANJEVANJA PODATKOV 21

3.1 HIERARHIČNI MODEL 21
3.2 MREŽNI MODEL 22
3.3 RELACIJSKI MODEL 23
3.4 OBJEKTNI MODEL 25
3.5 HIBRIDNI MODELI 26

4 KAKOVOST PODATKOV 29

4.1 POMEMBNOST KAKOVOSTI IN STROŠKI NEKAKOVOSTNIH PODATKOV 29
4.2 KAKOVOST PODATKOV IN RAZLIČNI INFORMACIJSKI SISTEMI 35
4.2.1 Kakovost podatkov in sistemi sprotne obdelave podatkov 35
4.2.2 Kakovost podatkov in skladišča podatkov 36
4.2.3 Kakovost podatkov in odkrivanje zakonitosti v podatkih 38
4.2.4 Kakovost podatkov in obvladovanje razmerij s kupci 41
4.3 RAZSEŽNOSTI KAKOVOSTI PODATKOV – MATRIKA KAKOVOSTI PODATKOV 44
4.3.1 Natančnost 47
4.3.2 Doslednost 49
4.3.3 Popolnost 51
4.3.4 Zaupanje v podatke 53
4.3.5 Pravočasnost 54
4.3.6 Ustreznost podatkov 56
4.3.7 Razumljivost podatkov 56
4.3.8 Dostopnost 56
4.3.9 Druge razsežnosti 58
4.3.10 Soodvisnost razsežnosti 58
4.4 NADGRAJENA MATRIKA KAKOVOSTI PODATKOV 59
4.4.1 Zbiranje in posredovanje podatkov 60
4.4.2 Nadzorovanje podatkov 61
4.4.3 Shranjevanje podatkov 61
4.4.4 Povezovanje podatkov 62
4.4.5 Analize podatkov 64

4.5 VZROKI NEKAKOVOSTNIH PODATKOV 64
4.5.1 Vnos podatkov 66
4.5.2 Staranje podatkov 67
4.5.3 ETL podatkov 68
4.5.4 Uporaba podatkov 70

5 ČIŠČENJE PODATKOV 71

5.1 MANJKAJOČE VREDNOSTI OZIROMA NEPOPOLNE VREDNOSTI 72
5.2 IZJEMNE VREDNOSTI 73
5.3 SAMODEJNO ČIŠČENJE PODATKOV 73
5.4 ROČNO ČIŠČENJE PODATKOV 75
5.5 ZDRUŽEK ČIŠČENJA PODATKOV 75

6 MERJENJE KAKOVOSTI PODATKOV – PREREZ PODATKOV 77

6.1 CILJI PREREZA PODATKOV 78
6.2 MODEL PREREZA PODATKOV 79
6.3 TEHNIKE PREREZA PODATKOV 80
6.3.1 Analiza stolpcev 81
6.3.2 Strukturna analiza podatkov 83
6.3.3 Analiza povezav med podatki 84

7 RAVNANJE S KAKOVOSTJO PODATKOV 85

7.1 TIQM 86
7.1.1 Informacija kot poslovni učinek 87
7.1.2 Informacijska mapa 88
7.2 STANDARDIZACIJA IN PROBLEMI S STANDARDI 90
7.3 RAVNANJE S KAKOVOSTJO PODATKOV S STALIŠČA REVIZIJE 91

8 POMEMBNEJŠI SKLEPI IZ TEORETIČNIH SPOZNANJ 93

8.1 SKLEP 1: LOČEVANJE IZRAZOV PODATEK IN INFORMACIJA 93
8.2 SKLEP 2: POMEMBNOST KAKOVOSTI PODATKOV ZA POSAMEZNO ZDRUŽBO 95
8.3 SKLEP 3: SUBJEKTIVNO IN OBJEKTIVNO MERJENJE KAKOVOSTI PODATKOV 96
8.4 SKLEP 4: OPREDELITEV KAKOVOSTI PODATKOV 100
8.5 SKLEP 5: IZBOR NAJPOMEMBNEJŠIH RAZSEŽNOSTI KAKOVOSTI PODATKOV 103
8.6 SKLEP 6: OPREDELITEV KAKOVOSTI IZVEDBE RELACIJSKEGA PODATKOVNEGA MODELA 104
8.7 SKLEP 7: PROUČEVANE LASTNOSTI RELACIJSKEGA PODATKOVNEGA MODELA IN RAZSEŽNOSTI KAKOVOSTI PODATKOV 109

9 RAZISKAVA 111

9.1 OPREDELITEV VIROV IN TIPOV PODATKOV RAZISKAVE 113
9.2 VIDIK RELACIJSKEGA PODATKOVNEGA MODELA 114
9.3 MODELIRANJE S SUBP – MS SQL 2005 118
9.3.1 Načrtovanje podatkovnega modela – scenarij 1: Vrtec 118
9.3.2 Načrtovanje podatkovnega modela – scenarij 2: Prodaja blaga 127
9.3.3 Razprava na podlagi modeliranja 134
9.4 PREIZKUŠANJE MODELA S STATISTIČNO ANALIZO 140
9.4.1 Vzorec združb 140
9.4.2 Osnovne opisne statistike in primerjava z mednarodnima raziskavama 144
9.4.3 Obdelava odgovorov o kakovosti podatkov 147

9.4.4 Obdelava odgovorov o kakovosti zbirke podatkov 151
9.4.5 Obdelava izvedenih trditev 153
9.4.6 Razprava na podlagi statistične analize in povzemanje drugih ugotovitev 162
9.5 POTRDITEV GLAVNE TEZE 165
10 SKLEP 167
STVARNO KAZALO 169
LITERATURA 171

Kazalo preglednic:

Preglednica 1: Glavna trditev in izvedene trditve 8
Preglednica 2: Prednosti in slabosti posamezne podatkovne strukture 27
Preglednica 3: Opredelitev stroškov nekakovostnih podatkov 32
Preglednica 4: Obvladovanje razmerij s kupci 44
Preglednica 5: Razsežnosti kakovosti podatkov 46
Preglednica 6: Primer sintaktične natančnosti in nenatančnosti 48
Preglednica 7: Prikaz semantične in strukturne nedoslednosti 50
Preglednica 8: Popolnost z vidika vrednosti NULL 52
Preglednica 9: Druge razsežnosti, kot sta jih opredelila Wang in Strongova 58
Preglednica 10: Primer manjkajočih vrednosti 73
Preglednica 11: Simboli informacijske mape 89
Preglednica 12: meta podatkovni model za informacijsko mapo 90
Preglednica 13: Primer 3NF, ki ni BCNF 107
Preglednica 14: Opredelitev razsežnosti kakovosti relacijske podatkovne sheme 109
Preglednica 15: Primerjava raziskovalnih metod 112
Preglednica 16: Proučevanje vpliva FIRPM na kakovost podatkovs pomočjo modeliranja 135
Preglednica 17: Izpolnjeni anketni vprašalniki glede na velikost združb in primerjava s populacijo ... 141
Preglednica 18: Kontingenčna preglednica – velikost združb ter kakovost podatkov 141
Preglednica 19: Povezanost med velikostjo združb in razsežnostmi kakovosti podatkov 142
Preglednica 20: Povezanost med velikostjo združb in firpm 142
Preglednica 21: Vzroki za nekakovost podatkov 146
Preglednica 22: Opisna statistika odgovorov o kakovosti podatkov 147
Preglednica 23: Izračun kakovosti podatkov v anketiranih slovenskih združbah 148
Preglednica 24: Koeficient korelacije med posledicami odločitev in razsežnostmi kakovosti podatkov 148
Preglednica 25: Regresijska analiza za odvisno spremenljivko »zaupanje v podatke« 149
Preglednica 26: Kontingenčna preglednica – velikost združb ter popravljanje podatkov 149
Preglednica 27: Koeficient korelacije med pomembnostjo pis in pomembnostjo zbirke podatkov 150
Preglednica 28: Opisna statistika odgovorov o kakovosti zbirke podatkov 151
Preglednica 29: Koeficienti korelacije med razsežnostmi kakovosti podatkov in starostjo pis 153
Preglednica 30: Značilnosti gruče 5 153
Preglednica 31: Koeficienti korelacije med razsežnostmi kakovosti podatkov in firpm 154
Preglednica 32: Spremenljivke v algoritmu Naive Bayes 155
Preglednica 33: Izračuni z algoritmom Naive Bayes 156
Preglednica 34: Regresijska analiza za odvisno spremenljivko natančnost 157
Preglednica 35: Regresijska analiza za odvisno spremenljivko doslednost 158
Preglednica 36: Regresijska analiza za odvisno spremenljivko popolnost 161
Preglednica 37: Preglednica rezultatov raziskave 165

Kazalo slik:

Slika 1: Vloga informacijskih tehnologij v organizaciji 2
Slika 2: Povečevanja napak v odvisnosti od vgrajenih omejitev v PIS 7
Slika 3: Primer hierarhične zbirke podatkov 22
Slika 4: Hierarhični in mrežni podatkovni model 23
Slika 5: Primer relacijske zbirke podatkov 24
Slika 6: Primer shranjevanja podatkov v relacijskem (levo) in objektnem modelu (desno) 26
Slika 7: Razdelitev stroškov nekakovostnih podatkov 34
Slika 8: Odvisnosti med doslednostjo in deklarativnimi omejitvami ter številom ponavljanj 51
Slika 9: Soodvisnost med različnimi razsežnostmi 59
Slika 10: nekakovost v odvisnosti od časa pri starajočih se podatkih 68
Slika 11: Gartnerjev kvadrant orodij za ravnanje s kakovostjo podatkov 74
Slika 12: Razširjeni entitetno-relacijski diagram z vidikom kakovosti 88
Slika 13: Primer izdelave informacijske mape 89
Slika 14: Matrika primerjave subjektivne in objektivne ocene 98
Slika 15: Opredelitev kakovosti podatkov s povezavo statične in dinamične opredelitve 101
Slika 16: Shema opredelitve kakovosti podatkov 102
Slika 17: Prikaz najpomebnejših razsežnosti različnih avtorjev 103
Slika 18: Primer zgoščene predstavitve podatkovne sheme 105
Slika 19: Smer morebitne povezanosti oziroma vpliva med firpm in kakovostjo podatkov 110
Slika 20: Odvisnost med kakovostjo podatkov in analizami 117
Slika 21: Razširjeni entitetno-relacijski diagram z vidikom kakovosti – scenarij 1 119
Slika 22: Podatkovni model, manjša doslednost – scenarij 1 124
Slika 23: Podatkovni model, večja doslednost – scenarij 1 126
Slika 24: Razširjeni entitetno-relacijski diagram z vidikom kakovosti – scenarij 2 127
Slika 25: Začetni Podatkovni model – scenarij 2 129
Slika 26: Podatkovni model po denormalizaciji – scenarij 2 130
Slika 27: Prikaz odvisnosti med glavnimi ključi in doslednostjo 135
Slika 28: Anketirane združbe po panogah dejavnosti 143
Slika 29: Anketirane združbe po urejenosti informatike 143
Slika 30: Podatki kot premoženje združbe 144
Slika 31: Zaznavanje kakovosti podatkov v združbah 145
Slika 32: Pobudniki razprav in ukrepov glede kakovosti podatkov 146
Slika 33: Prikaz odvisnosti s pomočjo algoritma Naive Bayes 155
Slika 34: Prikaz vrednosti za odvisno spremenljivko natančnost 157
Slika 35: Prikaz vrednosti za odvisno spremenljivko doslednost 159
Slika 36: Prikaz vrednosti za odvisno spremenljivko popolnost 162
Slika 37: Prikaz odvisnosti med deklarativnimi omejitvami in doslednostjo 164

1 UVOD

Kakovost podatkov je v novejšem času vedno bolj aktualna tema, ki je bila do zdaj nekoliko zanemarjena. Na začetku razvoja računalniško podprte informatike je bilo največ napora vloženega v razvoj strojne opreme, manj pozornosti pa je bilo namenjene razvoju programske opreme (Lesjak et al. 2006). Iznajdba računalnika, ki je omogočil tudi polet na Luno, je pomenila preskok na področju računalništva in posledično informatike. Polet na Luno je triumf združbe, ki je omogočila, da se desetletni trud več sto tisoč ljudi usmeri v ta projekt, toda prav tako je bil to triumf informatike, ki je bila poglavitno orodje tega uspeha (Mihelčič 1972, 76). Pomemben preskok na področju računalništva pomeni iznajdba mikroprocesorja, ki je nekoč okorne, velikanske naprave zmanjšal in jih posadil na mizo posameznika (Bellis 2006)

Računalniki so bili najprej razviti za potrebe vojske, in sicer za računanje trajektorij balističnih raket ter dešifriranje nemških in japonskih sporočil (Weik 1961). O shranjevanju podatkov in njihovi uporabi za poslovne odločitve takrat ni nihče razmišljal. Računalniki so bili preprosto predragi in premalo zmogljivi. Pogled na računalnike najbolje opisuje Watsonova izjava iz leta 1943, da je na svetu dovolj prostora le za pet računalnikov (Cerf in Navasky 1998, 230) ter Olsonova izjava iz leta 1977, da ni nobenega razloga, da bi kdor koli doma imel računalnik (Frazer 2007).

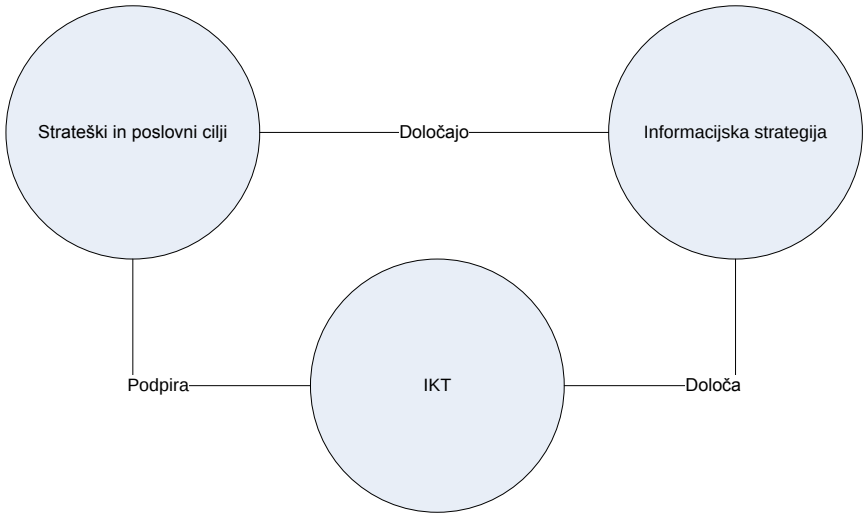
Doba shranjevanja podatkov se je začela z večjo dostopnostjo računalnikov in pocenitvijo nosilcev podatkov. Precejšnjo revolucijo je pomenil razvoj relacijskih podatkovnih zbirk. Osnove relacijskega podatkovnega modela je v 70. letih prejšnjega stoletja postavil Codd,[1] ki jih je leta 1985 dopolnil s svojimi 12 pravili, v 90. letih pa jim je dodal še šest pravil (Codd 1975, 1975, 1979, 1990). Predhodno so obstajale že hierarhične in mrežne zbirke podatkov, ki pa jih je danes relacijski model popolnoma izpodrinil. Teoretična osnova relacijskega modela zbirk podatkov je v začetku 80. let prejšnjega stoletja povzročila nastanek mnogih relacijskih sistemov za upravljanje zbirk podatkov (v nadaljevanju RSUBP) (angl. *Relational Database Management System* – RDBMS. Na začetku jih je bilo okrog 200 (Olson 2003, 37), danes pa prevladuje nekaj ponudnikov, na primer Oracle, MS SQL, DB2, MySQL, Sybase ter Postgre.

S pojavom ustrezne strojne in programske opreme se je začela doba zbiranja podatkov in njihova uporaba v vsakdanjem poslovnem življenju. Z nadaljnjim razvojem strojne opreme se je začela razvijati tudi namenska programska oprema za podporo odločanju. Nastali so različni

[1] Edgar Codd je objavil več člankov s področja relacijskega podatkovnega modela. Njegov delodajalec IBM je njegove zamisli najprej zanemarjal in šele leta 1978 se je multinacionalka odločila razviti uporabniško rešitev na podlagi njegovih zamisli, a jih je prehitel Lawrence J. Ellison, ki je pozneje ustanovil podjetje Oracle (Hafner, 2003).

sistemi za poslovno obveščanje, na primer sistemi za odkrivanje zakonitosti v podatkih (angl. *Data Mining*, v nadaljevanju SOZP), sistemi za obvladovanje razmerij s kupci, informacijski sistemi za strateški management itn. Največjo revolucijo v zadnjem obdobju pa je povzročil internet, ki je znatno spremenil način našega življenja in poslovanja ter omogočil storitve, o katerih pred desetletjem nismo niti razmišljali. Internet je v štirih letih dosegel 50 milijonov uporabnikov, kar do zdaj ni uspelo nobenemu še tako revolucionarnemu mediju (Greenstein in Feinman 2000, 6).

SLIKA 1: VLOGA INFORMACIJSKIH TEHNOLOGIJ V ORGANIZACIJI



Vir: Gunasekaran, Khalil & Rahman, Knowledge and information technology management: human and social perspective, 2003, str. 21.

Informacijska komunikacijska tehnologija (IKT, v nadaljevanju IT) ima izjemno pomembno vlogo pri poslovanju združbe. Management združbe določi poslovne cilje, poslovni cilji določajo informacijsko strategijo, slednja pa določa IKT (slika 1). Iz povedanega lahko sklepamo, da IT neposredno ali posredno vpliva na tekmovalno sposobnost združbe (Gunasekaran, Khalil in Rahman 2003)

Združbe oziroma natančneje podjetja so v kateri koli panogi pod vplivom petih tekmovalnih sil (Clarke 1994):

- nevarnosti vstopa novih podjetij (nevarnost se izraža predvsem z ovirami, ki (ne) preprečujejo vstop(-a) novih ponudnikov);
- pogajalske moči kupcev (določajo jo raven ozaveščenosti, stroški zamenjave dobavitelja);
- pogajalske moči dobaviteljev (na primer koncentracija dobaviteljev);
- nadomestnih poslovnih učinkov;
- tekmovanja med obstoječimi podjetji v panogi.

Te zunanje spremenljivke določajo privlačnost panoge, saj vplivajo na cene, stroške in naložbe, s tem pa tudi na ekonomsko uspešnost, ki jo panoga dosega (Pučko 1996).
Kako torej IT lahko vpliva na tekmovalno sposobnost podjetja z vidika Porterjevih dejavnikov? Cats-Baril in Thompson (2002) pravita:

- Z IT podjetje lahko zniža stroške razvoja novega poslovnega učinka, avtomatizira njegovo ustvarjanje in poveže dobavitelje s svojim sistemom, kar povišuje stroške vstopa drugih podjetij. Podjetja, ki bi hotela vstopiti na ta trg, bi morala biti vsaj tako dobra kot boljša obstoječa podjetja na trgu.
  Sklep: **IT vpliva na dejavnik »nevarnosti vstopa novih podjetij«.**
- IT na različne načine povečuje raven kupčeve zvestobe; na primer spletna knjigarna Amazon kupcem uspešno predlaga nakup na podlagi preteklih nakupov. Storitev je zelo uspešna, zato lahko rečemo, da IT vpliva na večjo zvestobo kupcev, poleg tega pa vpliva tudi na njihovo znanje, zaradi česar imajo slednji večjo pogajalsko moč (na primer kupec lahko hitro opravi primerjavo s tekmovalnimi poslovnimi učinki).
  Sklep: **IT vpliva na dejavnik »pogajalska moč kupcev«.**
- Na razvitih trgih podjetja največkrat tekmujejo s ceno. Obstajajo pa tudi druge možnosti, na primer znižanje stroškov, znatno večja produktivnost, dodajanje novih funkcij poslovnim učinkom. Z informacijsko tehnologijo (CAD/CAM) lahko hitro prilagodimo določene poslovne učinke in zamenjamo tekmovalno sposobnost (nizkocenovni poslovni učinek z novo funkcijo, ki jo tekmovalni poslovni učinki nimajo.)
  Sklep: **IT vpliva na dejavnik »nadomestni poslovni učinki«.**
- Z napredno informacijsko tehnologijo je veliko storitev postalo bolj prijaznih do uporabnika. Danes je letalsko vozovnico mogoče rezervirati prek spleta (EasyJet, RyanAir) brez posrednikov, zato so letalske vozovnice cenejše. Poleg tega pa so različni posredniki izgubili svojo pogajalsko moč.
  Sklep: **IT vpliva na dejavnik pogajalska »moč dobaviteljev«.**
- IT pa vpliva tudi na tekmovanje med obstoječimi podjetji v panogi.

Podatki so osnova IT. Računalniška informacijska revolucija in globalizacija sta povzročili, da so združbe začele zbirati podatke iz različnih virov. Gre predvsem za podatke o kupcih, dobaviteljih, tekmecih, poslovnih učinkih itn. Pozornost združb je bila na začetku usmerjena predvsem v zbiranje podatkov, nihče oziroma zelo malo združb pa se je dejansko ukvarjalo s kakovostjo podatkov. Zato je bilo samo še vprašanje časa, kdaj se bodo združbe v hudem tekmovalnem boju osredotočile na kakovost podatkov v smislu vzpostavitve mehanizma obvladovanja kakovosti podatkov (Bowen, Fuhrer in Guess 1998). Tudi na raziskovalnem področju je bila

kakovost podatkov dolgo časa zanemarjeno področje, saj so se raziskovalci posvečali predvsem umestitvi IS v združbe (glej npr. Mihelčič 1972).

Kakovost podatkov je danes izredno aktualno, pomembno in zelo hitro razvijajoče se področje. Če avtorja Batini in Scannapieca (2006, 1) navajata, da iskalno geslo »Data Quality« vrne približno tri milijone zadetkov, julija 2007 enako iskalno geslo vrne šest milijonov zadetkov. Pomembnost tega področja nakazujejo tudi ukrepi Evropske unije, na primer EUROSTAT na svojih spletnih straneh navaja zagotavljanje kakovostnih podatkov kot enega od glavnih ciljev delovanja te institucije. Od leta 1996 je vsako leto organizirana konferenca o kakovosti informacij (angl. *International Conference on Information Quality – ICIQ*) (MIT 2007), katere glavni organizator je tehnološki inštitut iz *Massachusettsa* (*Massachusetts Institute of Technology*, v nadaljevanju MIT). O razvoju in pomembnosti področja pričajo tudi prispevki na omenjeni konferenci, ki so vsako leto številčnejši in boljši.

Oblikoval se je tudi prvi podiplomski program na univerzi v Arkansasu (glej UALR 2007) z naslovom *Master of Science in Information Quality*, v pripravi pa je že tudi doktorski študijski program. Univerza *Dublin City* na svojih spletnih straneh objavlja dokončane doktorske disertacije in disertacije v nastajanju s področja kakovosti podatkov (glej University 2007). Decembra 2007 je bilo dokončanih deset disertacij, dve disertaciji pa sta še nastajali.[2] Ustanovljena je tudi nepridobitna mednarodna zveza za kakovost podatkov in informacij, katere namen je širiti zavest o pomembnosti kakovosti podatkov in informacij (glej IAIDQ 2007).

Kot dopolnilo programu, ki teče na MIT, in mednarodni konferenci o kakovosti podatkov je konec leta 2006 začela izhajati tudi revija *Journal of Data and Information Quality* (JDIQ) (glej JDIQ 2007), ki izhaja štirikrat na leto.

Pri pregledu svetovne literature na področju kakovosti podatkov izstopajo Don Ballou in Harry Pazer, ki sta začetnika področja kakovosti podatkov[3] (glej University 2007), Richard Wang, ki je direktor programa za zagotavljanje kakovosti podatkov na MIT, Thomas Redman, katerega vzdevek je The Data Doc, Larry English, ki je razvil metodo *Total Quality Data Management*, temelječo na metodi KAIZEN, Jack Olson z metodo prereza podatkov ter David Loshin.

Kljub izredni aktualnosti in zanimivosti področja je v svetovnem merilu mogoče zaslediti le eno raziskavo, in sicer raziskavo združbe *The Data Warehouse Institute* (v nadaljevanju TDWI) z naslovom »*Taking Data Quality to the Enterprise through Data Governance*«. Raziskava je poskušala ugotoviti vzroke, posledice nekakovosti podatkov ter koristi kakovostnih podatkov v združbah v ZDA. V raziskavi sta bili uporabljeni kvalitativna (intervju) in kvantitativna

[2] Upoštevane so disertacije, ki so jih avtorji prijavili na spletno stran.
[3] Po njiju se imenuje tudi nagrada za doktorsko disertacijo s področja kakovosti podatkov.

metodologija (spletna anketa), pri čemer je TDWI ciljala predvsem na analitike kakovosti podatkov, svetovalce za informatiko ter dobavitelje programske opreme. Vabilo za sodelovanje v raziskavi je TDWI poslala udeležencem po elektronski pošti, poleg tega pa je vabilo objavila tudi na več spletnih straneh. Vabilu se je odzvalo 399 združb, ki so navedle naslednje vzroke nekakovostnih podatkov (od najbolj pogostega do najmanj pogostega) (Russom 2006)::

- nepopolne opredelitve posameznih entitet,
- napačen vnos podatkov zaposlenih,
- različni prehodi in obdelave podatkov,
- različna pričakovanja uporabnikov,
- zunanji podatki,
- napačen vnos podatkov strank,
- sistemske napake,
- spremembe vhodnih postopkov,
- drugo.

Vzroke za nekakovostne podatke lahko razdelimo v tri osnovne skupine:

- organizacija poslovanja,
- zgradbo uporabniške rešitve,
- okvare v delovanju programske in strojne opreme.

Groznik in Vičič (2006) ugotavljata, da imajo projekti v današnjem času ključno vlogo za uspeh vsake združbe in tudi ukvarjanje s kakovostjo podatkov se po navadi začne kot projekt (Sanna 2006), na primer kot izdelava skladišča podatkov in nadgraditev z OLAP. Če takšen projekt postane uspešen, lahko to spodbudi tudi druge, da se vključijo v proces izboljšanja kakovosti podatkov. Pomembno je, da se s problematiko začne ukvarjati tudi management združb; na primer z uvedbo različnih notranjih predpisov poslovanja, ki natančno določijo vloge posameznika. Olson (2003) navaja dve metodi reševanja navedene problematike:

- od znotraj navzven (angl. *inside-out*) – preprečevanje nekakovostnih podatkov in
- od zunaj navznoter (angl. *outside-in*) – odpravljanje nekakovostnih podatkov.

K temu bi lahko dodali še metodi od zgoraj navzdol (angl. *top-bottom*) in od spodaj navzgor (angl. *bottom-up*). Prva predpostavlja, da strateški management ni ozaveščen o problematiki, zato je pobudnik izvedbeni management. Druga metoda pa predpostavlja, da je management združb ozaveščen in vzpostavi ustrezno organizacijsko sestavo, ki vodi do boljše kakovosti podatkov.

Z razmahom zbirk podatkov, predvsem analitičnih, se je pojavila tudi potreba po novem poklicu, in sicer analitiku kakovosti podatkov. Pirceova (2003) pripisuje to predvsem neuspehu številnih analitičnih projektov. Glavna naloga analitika kakovosti podatkov naj bi bila izpolnjevanje zahtev managementa s pomočjo podatkov, urejenih v skladiščih podatkov. Podrobnejše zadolžitve analitika kakovosti podatkov so navedene v knjigi *Managing the Data Warehouse*, in sicer (Inmon, Welch in Glassey 1997):

- pregled natančnosti podatkov, prenesenih v skladišče podatkov,
- predlaganje izboljšav v procesu zajemanja podatkov z namenom izboljšati kakovost podatkov v skladiščih podatkov,
- pregled povezav med podatki (v nadaljevanju referenčnih integritet) v skladišču podatkov,
- pregled zgodovinske usklajenosti podatkov v skladišču podatkov.

Avtorji sicer navajajo zadolžitve oziroma odgovornosti analitika, ne posvečajo pa se potrebnim sposobnostim, ki naj bi jih analitik imel. To vrzel poskušajo zapolniti Chung, Fisher in Wang (2002), ki priporočajo naslednje sposobnosti:

- tehnične sposobnosti – znanje SOZP, znanje statistike, poizvedovalnega jezika (SQL) ter znanje programskih jezikov,
- ekipne sposobnosti – sposobnost komuniciranja in usklajevanja med končnimi uporabniki ter managementom združb in drugimi udeleženci v poslovnih procesih,
- sposobnost razlage – sposobnost zaznavanja vpliva kakovosti/nekakovosti podatkov na poslovanje, zaznavanje napak v modelih podatkov ter sposobnost izvedbe stroški – koristi analize pri ukrepih za povečanje kakovosti podatkov.

Vsekakor je področje kakovosti podatkov zelo zanimivo in obetavno ter ponuja veliko možnosti za izboljšave v združbah v smislu znižanja stroškov, prihrankov oziroma povečanje dobička ali zmanjšanja izgube. Poleg tega ponuja tudi priložnost za udejstvovanje mladih strokovnjakov, ki jih zanima interdisciplinarnost področja.

V letu 2005 je bila opravljena že omenjena raziskava TDWI (Russom 2006), ki je poleg ocenjevanja stroškov nekakovostnih podatkov ugotavljala tudi vzroke za njihov nastanek.

Stroški in vzroki za nastanek nekakovostnih podatkov so bili motivacija za razmišljanje o vplivu fizične izvedbe relacijskega podatkovnega modela (v nadaljevanju FIRPM) na kakovost podatkov. S fizično izvedbo se predpostavlja prisotnost glavnih ključev, tujih ključev, uporabo ustreznih podatkovnih tipov, normalizacije ter uporabo deklarativnih omejitev.

SLIKA 2: POVEČEVANJA NAPAK V ODVISNOSTI OD VGRAJENIH OMEJITEV V PIS

Na sliki 2 lahko vidimo, kako se možnost za nastanek nekakovostnih podatkov povečuje z oddaljevanjem od prve omejitve, tj. tipom polja v zbirki podatkov. Z ustreznim FIRPM lahko ploščino, ki pomeni količino nekakovostnih podatkov, zelo omejimo, kar izvedena razsikava predstavljena v knjigi poskuša tudi dokazati.

Cilj raziskave je pokazati odvisnost med FIRPM in kakovostjo podatkov, torej:

- povezanost med glavnimi ključi in kakovostjo podatkov,
- povezanost med tujimi ključi in kakovostjo podatkov,
- povezanost med podatkovnimi tipi in kakovostjo podatkov,
- povezanost med stopnjo normalizacije/denormalizacije in kakovostjo podatkov,
- povezanost med deklarativnimi omejitvami in kakovostjo podatkov.

1.1 Hipoteze raziskave

Predmet mojega zanimanja sta prvi dve skupini vzrokov za nastanek nekakovostnih podatkov, navedeni v raziskavi TDWI (Russom 2006) in sicer da sta najpogostejša vzroka za nastanek slabih podatkov dva: prvi je ta, da so entitete nepopolno opredeljenje, drugi pa je ta, da zaposleni vnašajo napačne podatke. To bi z ustreznim podatkovnim modelom lahko omejili. Svoje raziskovanje sem usmeril na relacijski podatkovni model, ki je trenutno najbolj razširjen in popularen (Gilfillan 2002). Glavni namen moje knjige je torej proučiti, **kako podatkovni model vpliva na kakovost podatkov v PIS**. Podatkovni model je najpogosteje temelj uporabniških rešitev in kakovostna zasnova je po mojem mnenju izjemno pomembna za kakovost podatkov. Glavna trditev knjige pred vami se torej glasi:

Podatkovni model vpliva na kakovost podatkov v PIS. Pri tej trditvi me zanima, kako pravilne opredelitve podatkovnih tipov, glavnih ključev, tujih ključev, omejitev na posameznih poljih ter raven normalizacije/denormalizacije vplivajo na kakovost podatkov. Ker je glavna trditev dokaj

splošna, sem jo razčlenil na izvedene trditve, ki so bolj ozko usmerjene in katerih potrditev ali zavrnitev mi bo pomagala potrditi ali ovreči glavno trditev.

Izvedene trditve:

Glavni ključi (angl. *primary key*) podatkovnih entitet vplivajo na kakovost podatkov v PIS. Glavni ključi so eden od temeljev podatkovnega modela, saj enolično določajo zapis znotraj posamezne preglednice. Odsotnost ali neustrezni glavni ključi posledično lahko vplivajo na slabšo kakovost podatkov, hkrati pa lahko imajo hude posledice v delovanju uporabniške rešitve, vendar slednje ni predmet te knjige (glavni ključ, sestavljen iz mnogo polj, potrebuje več virov za delovanje kot krajši glavni ključ).

Tuji ključi (angl. *foreign key*) podatkovnih entitet vplivajo na kakovost podatkov v PIS. Tuji ključi preprečujejo pojavljanje »sirot« v zbirki podatkov v povezanih preglednicah. Tako ne moremo vnesti zapisa v odvisno preglednico, če določen atribut tega zapisa ni vnesen v glavno/primarno preglednico. Prav tako ne moremo brisati zapisa v glavni/primarni preglednici, če ima ta zapis podrejene zapise v odvisnih preglednicah. To zelo zmanjša možnost napak v zbirki podatkov (na primer pojavljanje vrstic računa, ki pa ne bi imele pripadajoče glave računa).

PREGLEDNICA 1: GLAVNA TRDITEV IN IZVEDENE TRDITVE

Glavna trditev	**Podatkovni model vpliva na kakovost podatkov v PIS.**
1. izvedena trditev	*Glavni ključi podatkovnih entitet vplivajo na kakovost podatkov v PIS.*
2. izvedena trditev	*Tuji ključi podatkovnih entitet vplivajo na kakovost podatkov v PIS.*
3. izvedena trditev	*Podatkovni tipi polj vplivajo na kakovost podatkov v PIS.*
4. izvedena trditev	*Deklarativne omejitve polj vplivajo na kakovost podatkov v PIS.*
5. izvedena trditev	*Normalizacija v zbirki podatkov vpliva na kakovost podatkov v PIS.*

Podatkovni tipi polj vplivajo na kakovost podatkov v PIS. Kakovost podatkov začnemo obvladovati v zbirki podatkov, natančneje v posameznih poljih oziroma z opredelitvijo polja. Z oddaljevanjem od polja je pravila mogoče lažje zaobiti, možnost za napake pa se poveča. Naj to pojasnim s primerom polja, ki vsebuje podatek o zemljepisni širini. Vrednost 38 stopinj in 40 minut bi lahko vpisali v katero koli tekstovno polje, a če to polje opredelimo kot številčno polje, se možnost za napake zelo zmanjša. Najboljše bi bilo opredeliti polje, ki bi zahtevalo vnos v točno določeni obliki, kar pa zahteva več truda in povišanje stroškov, zato se za tako strategijo po navadi odloči zelo malo ponudnikov informacijskih sistemov. Če pa je pravilo določeno na ravni uporabniške rešitve, je možnost za napake že večja, še posebno če ima uporabnik določene rešitve dovolj znanja in dostopa do podatkov tudi mimo uporabniške rešitve.

Deklarativne omejitve polj vplivajo na kakovost podatkov v PIS. Deklarativne omejitve polj preprečujejo vnos neželenih vrednosti v določeno polje in so s tega vidika nadgradnja podatkovnih tipov. Če s podatkovnim tipom določimo, ali v polje lahko vnesemo na primer samo številke (cela števila, decimalna števila), z deklarativno omejitvijo še podrobneje določimo sprejemljive vrednosti (na primer polje ocena bi tako imelo vrednosti v razponu od 1 do 10).

Normalizacija v zbirki podatkov vpliva na kakovost podatkov v PIS. Normalizacija je eden od temeljnih postopkov pri načrtovanju podatkovnega modela v sistemih sprotne obdelave podatkov (v nadaljevanju SSOP). Nenormaliziran model podatkov povzroča podvajanje podatkov, kar oteži postopek vzdrževanja podatkov. Če imamo na primer v zbirki podatkov naslov stranke zapisan v več preglednicah in stranka sporoči spremenjen naslov, ki ga pozabimo spremeniti na določenem mestu, postanejo podatki posledično nedosledni.

1.2 Sestava knjige

Kniga vsebuje teoretično-analitični ter izkustveni del. Teoretično-analitični del vsebuje pregled pomembnejše svetovne literature, ki se nanaša na izbrano problematiko. Pri tem uporabim metodo opisovanja. Ta je primerna, kadar gre za opis danih dejstev, ki jih kaže upoštevati pri razvijanju spoznanj in modela za razrešitev izbranega problema raziskovanja. Z metodo opisovanja opredelim dejavnike, ki vplivajo na kakovost podatkov, ter utemeljeno sprejmem ali zavrnem dana dejstva. V teoretičnem delu se osredotočam na:

- teoretični del o zbirkah podatkov, ki so temelj shranjevanja podatkov. V tem delu knjige se osredotočam tudi na glavne pojme, ki so vezani na zbirke podatkov;
- teoretični del o kakovosti podatkov. Za ponazoritev procesa doseganja kakovosti podatkov uporabim metodo *Total Information Quality Management* (v nadaljevanju TIQM). TIQM sem izbral zato, ker je usmerjen predvsem na načrtovanje in spremembe procesov, ki so vzrok nekakovosti v združbi z vidika stranke, izvede pa se lahko kot projekt *Six Sigma*, kar pa ni pogoj. Poleg metode TIQM prikažem tudi uporabo IT presoje v povezavi s kakovostjo podatkov.

Po teoretično-analitičnem delu se posvetim izkustvenemu delu, kjer uporabim kvalitativno in kvantitativno metodo. Za dokazovanje trditev sem izvedel dva scenarija na SUBP MS SQL 2005, ki prikazujeta delovanje proučevanih dejavnikov (izvedenih trditev) v praksi in sta bila osnova za model, kateri dejavniki FIRPM vplivajo na kakovost podatkov. S kvantitativno metodo sem zbral in obdelal podatke o uporabnikih PIS ter njihovem podatkovnem modelu.

Anketiranje se je nanašalo na:

- slovenske združbe kot uporabnike PIS – zajel sem manjše, srednje velike in večje združbe različnih panog in velikosti iz vseh statističnih regij. V vzorec niso bili vključeni samostojni podjetniki ter združbe, ki niso oddale zaključnega računa. Vzorec je bil nepristranski in ustvarjen naključno;
- ponudnike PIS, kjer sem dobil manjkajoče podatke o podatkovnih modelih za potrditev ali zavrnitev izvedenih trditev.

1.3 Kaj je že dognano in kaj ni?

Pri pregledu svetovne literature nisem naletel na nobeno raziskavo, ki bi z izkustveno metodo potrjevala oziroma zavračala moje trditve. Vendar problematika ni popolnoma nova. Zelo blizu področja moje knjige so avtorji članka *A formal definition of data quality problems* iz leta 2005 (Oliveira, Rodrigues in Henriques 2006). V svojem članku navajajo negativne posledice slabega relacijskega podatkovnega modela za kakovost podatkov. Ne ukvarjajo se z opredelitvijo kakovosti relacijskega modela, ampak samo s posledicami nekakovosti. Nekakovost proučujejo na ravni atributa, entitete, zbirke in več zbirk.

Področja kakovosti podatkov in SUBP sta se dotaknila tudi Naumann in Rothova (2004). Avtorja v svojem delu proučujeta relacijski SUBP ter njegov vpliv na kakovost podatkov. Največ pozornosti sta posvetila IBM-ovemu sistemu DB2, saj Rothova prihaja prav iz te združbe. Prav zaradi tega je morda članek v določenih delih pristranski, vendar to ne vpliva na njun sklep, da sodobni relacijski SUBP vpliva na kakovost podatkov.

Še najbliže temi knjige je Scott Ambler (2003), ki v svoji knjigi iz leta 2003 trdi, da preizkušanje relacijske zbirke podatkov pozitivno vpliva na njeno kakovost. Avtor priporoča preizkušanje postopkov, sprožilcev, pogledov, referenčnih integritet ter deklarativnih omejitev. Avtor s preizkušanjem misli na dejansko preizkušanje sheme, vsebine in funkcionalnosti zbirke podatkov. V svoji raziskavi iz septembra 2006 (Ambler 2006) pa se je dotaknil tudi povezave med preizkušanjem zbirke podatkov v postopku razvoja podatkovnega modela in kakovosti podatkov. Ugotovil je, da pogostejši preizkusi vodijo h kakovostnejšim podatkom. V svojih delih ni izpostavil podatkovnega tipa ter stopnje normalizacije/denormalizacije kot dejavnikov, ki bi ju bilo treba preizkusiti, ker bi imela pomemben vpliv na kakovost podatkov. Tukaj knjiga nadgrajuje njegova spoznanja. Poleg tega Ambler trditev v svojem delu iz leta 2003 ni izkustveno potrdil, ampak jih je postavil na podlagi lastnih izkušenj. Tudi sam lahko na podlagi izkušenj potrdim, da dejavniki, našteti v mojih trditvah, vplivajo na kakovost podatkov, vendar je treba trditve tudi izkustveno potrditi.

Povezavo med relacijsko podatkovno shemo in kakovostjo podatkov poleg Amblerja (2003, 2006) obravnavajo tudi Redman (1996) ter Batini in Scannapieca (2006). Omenjeni avtorji svoje trditve prav tako postavljajo na podlagi izkušenj, in ne na podlagi izkustvenih raziskav.

Knjiga obogati znanje oz. spoznanja na treh področjih:

- enotna opredelitev kakovosti podatkov – v knjigi so prikazane tri opredelitve kakovosti podatkov, ki jih je mogoče zaslediti v literaturi, na žalost pa niso povezane. Obstajata statična in dinamična opredelitev kakovosti podatkov, poleg tega pa je posebej obravnavan namen uporabe. Vse tri metode so v disertaciji združene in opisane z matematično enačbo. Matematična enačba je uporabljena tudi za izračun kakovosti podatkov v anketiranih združbah, s čimer se pokaže praktična vrednost enotne opredelitve kakovosti podatkov;
- izpeljava teoretičnega modela povezanosti FIRPM in kakovosti podatkov ter izkustvena preverba modela s pomočjo podatkov, dobljenih z anketiranjem – celovit teoretični model ter preverba z izkustveno metodo sta nadgraditev spoznanj, ki v literaturi že obstajajo. V literaturi je mogoče zaslediti določene elementarne povezave med FIRPM in kakovostjo podatkov, ki jih knjiga z modeliranjem poveže v smiselno celoto, izkustvena metoda pa to celoto opiše s funkcijo na podlagi podatkov, dobljenih z anketiranjem;
- uporaba izkustvene metode pa dodatno doprinese k znanosti pri ugotavljanju vloge FIRPM kot celote za kakovost podatkov – v literaturi je navedeno mnogo vzrokov za nekakovostne podatke, v knjigi pa je izpostavljen problem FIRPM. Knjiga torej da tudi odgovor na to, koliko lahko z ustreznim FIRPM vplivamo na kakovost podatkov. Ugotovitev je pomembna za nadaljnje raziskovanje, saj ugotovitev usmerja prihodnjo pozornost na tiste dejavnike, ki najbolj vplivajo na kakovost podatkov.

Knjiga skuša ugotoviti povezavo med glavnimi ključi, tujimi ključi, podatkovnimi tipi, deklarativnimi omejitvami, stopnjo normalizacije/denormalizacije ter kakovostjo podatkov. Zatorej je nadgraditev ter razširitev že znanih dejstev in spoznanj. Upam, da bo potrditev glavne trditve pripomogla k temu, da se bodo ponudniki PIS še bolj zavedali pomembnosti izdelave podatkovnega modela. Vse prevečkrat se po mojem védenju oziroma izkušnjah dogaja, da se načrtovanja FIRPM lotevajo posamezniki, ki nimajo zadosti znanja. Posledično je vsaka gradnja poslovne logike in uporabniškega vmesnika manj zanesljiva, ker stoji na trhlih temeljih. Načrtovalci podatkovnega modela pa se morajo zavedati še nečesa, in sicer da kakovosten podatkovni model navadno »preživi« več zamenjav poslovne logike in uporabniškega vmesnika.[4]

[4] Podjetji Mit Inženiring in SAOP Računalništvo, na primer, sta svoja podatkovna modela podedovali še iz časov Clipperja ter COBOL-a.

2 PODATKI, PROBLEMI IN ODLOČITVE

Da bi bolje razumeli vlogo podatkov, je v knjigi osnovam o podatkih namenjeno kar nekaj prostora, in sicer poglavje o odločanju ter o zaznavanju stvarnega sveta in prenosu na abstraktno raven, kjer potrebujemo podatke oziroma informacije, ter poglavje o zbirkah podatkov kot osnovni hrambi podatkov. Najprej pa si poglejmo, kako ljudje zaznavamo svet okoli sebe.

2.1 Znanje in modeli

Ljudje od nekdaj želimo vplivati na svet in pravzaprav nam to tudi uspeva. Poleg vpliva na svet pa imamo ljudje še eno veliko posebnost, posebnost modelirati stvarni svet in zaznavati vzorce v stvarnem svetu. In prav slednje je v veliki meri pritegovalo ljudi skozi vso zgodovino z namenom zgraditi modele stvarnega sveta.

Dogodki

V stvarnem svetu se ves čas nekaj dogaja: list z drevesa pade na tla, sonce vzide in zaide … Če je začetek nekega dogodka lahko enostaven, na primer skala se odtrga s pobočja, je posledic lahko veliko. Skala, ki se je odtrgala s pobočja, je zadela drugo skalo, ki se je tudi odtrgala in sprožila verigo dogodkov, ki lahko imajo hude posledice. Po nekem dogodku stanje ni več takšno, kot je bilo pred tem dogodkom. Pomembno pa je vedeti, da o dogodku lahko govorimo samo v povezavi z nečim. Dogodek kot osamljena enota ne more obstajati, vedno se mora navezovati na »nekaj drugega« (Pyle 2003) in nekaj drugega je objekt.

Objekti

Če je torej dogodek nekaj, kar se zgodi, je objekt nekaj, čemur se nekaj dogaja. Za objekt lahko rečemo, da je skupina določenih značilnosti, ki predstavljajo objekt, in opredeliti objekt je navadno zelo pomembno in časovno zelo obsežno opravilo. Tako Pyle (2003) govori o svoji izkušnji, ko je za opredelitev objekta stranka potreboval kar dva tedna. Pri opredelitvi objektov je pomembno upoštevati dejstvo, da začetna togost pri opredelitvi objektov lahko povzroči visoke stroške poznejše vnovične opredelitve objekta.

Zaznavanje

Zaznavanje pri ljudeh je največkrat omejeno z znanjem, ki ga imamo o določenih razmerah. Ljudje z določenim znanjem lahko imajo popolnoma drugačen pogled na svet kot ljudje z

drugačnim znanjem. In da presežemo omenjeni okvir (znanje), je potrebno veliko truda in naporov. Vendar nam prav ta »presežek« omogoča popolnoma nov pogled na dogodke v primerjavi s preteklostjo.

Podatki

Kaj je torej podatek? Za podatek lahko ugotovimo, da je niz znakov, ki pojasnjuje pojave. Je torej opredmetenje oziroma predstavitev dejstev, pojmov, predstav in znanja. Podatek je na neki način ugotovljeno in zapisano dejstvo, ki izraža neko stanje oziroma dogodek v sistemu ali njegovem okolju. Pravzaprav se v njih izražajo dogodki in stanja stvarnega ali miselnega sveta (Lesjak et al. 2006). Podatki shranjujejo vedenje o pojavih, ki jih trenutno znanje opredeljuje kot pomembne pojave.

Ureditve

Podatki torej shranjujejo vrednosti o objektih, ki nas zanimajo, in na tej osnovi se lahko začne gradnja ureditev. Ureditve so izredno pomembne za prikazovanje znanja, vendar je njihova statičnost velika ovira za bolj razširjeno uporabo. S statičnostjo se razume statična ponazoritev povezav med objekti.

2.2 Odločitvene razmere in odločitve

Herrenstein je leta 1961 izjavil, da je vsakršno vedenje pravzaprav odločanje (Steglich 2003, 29). Večina avtorjev opredeljuje odločanje oziroma odločitvene razmere kot problem, v katerem se odločevalec odloča med najmanj dvema različnima možnostma, med katerima izbere eno. Predpostavlja se, da odločevalec izbere posamezno možnost na podlagi prednosti in slabosti vsake možnosti. Odločitvene razmere so torej celota (Steglich 2003):

- pojava, v zvezi s katerim moramo nekaj ukreniti;
- odločevalca;
- znanja oziroma informacij, ki opredeljujejo pojav ter prostor mogočih rešitev.

2.2.1 Odločitveni okvir

Osnovni pogoj, da odločevalec sploh začne proces odločanja, so naslednji dejavniki v odločitvenih razmerah (Lesjak et al. 2006, 5):

- Obstajati mora dovolj velik razkorak med želenim (ciljnim) in problemskim stanjem. Če smo blizu tega, kar želimo, sprememba, ki bi odpravila odstopanja, ni potrebna.

- Odločevalec mora biti motiviran, da bo zmanjšal razkorak. Razkorak se mora, čeprav velik, nanašati na cilj, ki je za odločevalca pomemben.
- Odločevalec mora verjeti, da je razkorak mogoče odpraviti. Če je videti, da razkoraka ni mogoče odpraviti, se ga odločevalec ne bo lotil.
- Odločevalec mora seveda imeti tudi možnost, da to lahko naredi.

Lesjak in soavtorji (Lesjak et al. 2006) ugotavljajo, da posamezne odločitvene razmere opredeljujejo naslednji dejavniki:

- **negotovost razmer**, iz katerih odločitvene razmere izhajajo. Če so odločitvene razmere negotove, to pomeni, da je opredeljivost odločitvenih razmer slabša, prevladuje odločanje na podlagi izkušenj in intuicije. Če so odločitvene razmere opredeljene, pa prevladuje analitičen način odločanja;
- **pomembnost odločitve**. S pomembnostjo odločitve razumemo posledice akcij, ki so posledica naših odločitev. In kot pri negotovosti razmer tudi pomembnost odločitve opredeljuje različne načine odločanja;
- **časovna stiska** odločevalca oziroma čas, ki je na voljo za izvedbo odločitve. Če je časovna stiska velika, odločevalci niti ne zmorejo zbrati dovolj informacij ali proučiti številnih alternativ. V časovni stiski analitičen način odločanja ni primeren; bolj ustreza emocionalno-intuitiven način odločanja. Čas torej pomembno vpliva na način odločanja.

Proces odločanja lahko členimo na naslednje faze (Timmermans 1991):

- spoznavanje odločitvenih razmer (spoznavanje problema),
- členitev problema,
- zbiranje znanja,
- vrednotenje različnih možnosti,
- odločitev.

Pomembno je poudariti, da v praksi navedeni koraki niso tako jasno vidni in se mnogokrat prepletajo. V prvem koraku, torej koraku spoznavanja odločitvene situacije, spoznavamo okolje in pogoje za izvedbo procesa odločanja. Ta korak vsebuje dejavnosti spoznavanja in opredeljevanja odločitvenih razmer, ki je dejansko odkrivanje razlike med obstoječim in želenim stanjem. Obstoječe stanje primerjamo z želenim stanjem, opredelimo razkorak in razlike ovrednotimo. V okviru tega koraka je treba razumeti in natančno opredeliti odločitvene razmere ter tako ustvariti možnosti za snovanje različnih alternativ.

Členitev problema pomeni, da v procesu iskanja možnih rešitev problema določene možnosti že zavrnemo kot neprimerne. Po členitvi problema sledi zbiranje znanja. Izpostaviti je treba, da je

pri iskanju znanja priporočljiv ekonomski vidik zbiranja znanja, to pomeni poiskati znanje, ki obeta največ možnosti pri razreševanju odločitvenih razmer in ki izključuje potrebo po nadaljnjem iskanju znanja.[5] Zbiranje znanja pravzaprav pomeni iskanje različnih možnosti, med katerimi bo odločevalec izbiral.

V okviru koraka vrednotenja odločitev moramo opredeliti in razumeti prednosti, slabosti ter stroške vsake od zasnovanih odločitev. Ugotoviti moramo, kako te prednosti vplivajo na cilje odločanja in s kakšnimi posledicami. Vrednotenje odločitev je po Steiglichovem (2003) mnenju izkustveno (empirično) težko ločiti od predhodnega koraka, torej koraka kopičenja znanja. Po njegovem mnenju je bistvena razlika glede na zorni kot pogleda na možnosti (okvir posamezne možnosti proti vsem možnostim). Na koncu se odločimo, tj. izberemo najboljšo možnost med naborom možnosti. Pomembno pa je povedati, da posamezniki pogosto sklepajo, da je njihov osebni pogled na svet pravilen in da ga delijo z večino drugih. V odločitvenih razmerah posamezniki pogosto predpostavljajo, da če vsi vidijo dejstva tako kot oni, se bo vsakdo od njih enako ali podobno odločil. Ta predpostavka ni pravilna, ker ne upošteva vloge osebnosti v procesu odločanja in v zvezi s tem različnih načinov odločanja.

2.2.2 Urejenost odločitvenega procesa

Z besedo urejenost ponazarjamo lastnost odločitvenih razmer, in sicer opredeljujemo možnost vključitve informacijske tehnologije v proces odločanja. Za urejene odločitvene razmere velja, da natančno vemo, kako bodo potekale, natančno poznamo postopke odločanja. Druga skrajnost so neurejene odločitvene razmere, kjer pravila odločanja niso natančno znana. Urejene in neurejene odločitvene razmere pomenijo dve skrajnosti, med katerima obstajajo bolj ali manj urejene odločitvene razmere. Pri njih lahko predvidimo le posamezne dele odločitvenega procesa. Za druge dele pa tega ne moremo storiti. Urejene odločitve so torej odločitve z visoko stopnjo določljivosti oziroma programabilne odločitve. Neurejene odločitve pa so odločitve z nizko ravnjo določljivosti oziroma neprogramabilne odločitve. Lesjak (2006, 8) ugotavlja, da se je računalnik večinoma uporabljal le v tistem koraku procesa odločanja, ki ga označujemo kot spoznavanje odločitvenih razmer oziroma problema. Odločevalcu je zagotavljal vrsto podatkov, na podlagi katerih si je ta lahko ustvaril sliko (preteklega) dogajanja. Hkrati s temi prizadevanji pa si je znanstvena disciplina, ki jo pri nas označujemo kot operacijske raziskave, prizadevala z računalniško zasnovanimi kvantitativnimi modeli podpreti tisti korak odločitvenega procesa, ki ga označujemo kot odločanje v ožjem pomenu (izbiro možnosti). Ta korak procesa odločanja poskušajo računalniško podpreti tudi druge discipline, med katerimi izstopa umetna inteligenca. Hkrati s povečevanjem vloge računalnikov v odločitvenem procesu se je pojavila tudi želja po mislečem računalniku, tj. da bi računalnik razmišljal kot človek. Veliko je bilo razprav, ali je kaj

[5] Pri kopičenju znanja je zelo pomembno tudi »razmišljanje na glas« (glej Ranyard, Crozier in Svenson 1997)

takega sploh mogoče, vendar to presega okvir knjige. Dejstvo pa je, da sta človek in računalnik v neki zvezi, ali da je računalnik podrejen človeku, ali da računalnik lahko podpira posamezne človekove dejavnosti, ali pa podpira povezovanje odločevalcev v skupine.

2.2.3 Vrsta odločanja

Mnogo teorij govori o tem, na kakšen način se odločamo. Večina strokovnjakov opredeljuje tri načine odločanja (Lesjak et al. 2006, 10):

- racionalno-analitičen način,
- intuitivno-čustven način,
- vedenjsko-preudaren način.

Racionalno-analitičen način odločanja utemeljuje najstarejša odločitvena teorija, ki predpisuje racionalen, zavesten, sistematičen in analitičen način. Trdi, da je odločevalec samostojen in neodvisen dejavnik, katerega vedenje ni samo inteligentno, ampak prav tako racionalno. Odločitev je izbira, ki jo odločevalec izvede z namenom, da maksimira prednosti pri polnem poznavanju in zavedanju vseh razpoložljivih možnosti. Odločevalec torej prouči vse možnosti, prav tako pa tudi njihove posledice, jih razvrsti po pomembnosti in izbere možnost, ki zagotovi najboljše doseganje cilja. Ta način odločanja je bil pogosto kritiziran, ker:

- je odločevalec le redko neodvisen in samostojen dejavnik, navadno je le del zapletenih odločitvenih razmer;
- odločevalec ne pozna in niti ni zmožen proučiti vseh možnosti ali pa spoznati vseh posledic odločitve;
- odločevalec se ne odloča samo na osnovi maksimalnega doseganja ciljev, ampak tudi na osnovi na videz neracionalnih razlogov.

Intuitivno-emocionalen način odločanja je nasprotje racionalnega in trdi, da odločevalec daje prednost navadi ali izkušnjam, občutjem, instinktu, pri tem pa uporablja nezaveden miselni proces. Te procese lahko spodbudimo z »viharjenjem« možganov, ustvarjalno usmerjenostjo in ustvarjalnim soočanjem. Intuitiven odločevalec obravnava številne možnosti, tako da prehaja iz enega koraka v analizi ali iskanju v drug korak in nazaj. Nekateri od teh, ki dajejo prednost intuiciji, poudarjajo, da ta v mnogih primerih lahko vodi k boljšim odločitvam kot razne tehnike optimiranja.

Tisti, ki temu načinu odločanja nasprotujejo, trdijo, da odločevalec ne uporablja vseh sodobnih pripomočkov, ki so na voljo za odločanje (na primer operacijske raziskave, simulacije, sistemi za podporo odločanja ipd.), da racionalen način zagotavlja ustrezno pozornost posledicam odločitev, še preden so storjene velike napake, kar pa pri tem načinu odločanja ni v navadi.

Vedenjsko-preudaren način odločanja pa od odločevalca zahteva, da mora premisliti (in upoštevati) številne pritiske ljudi, na katere vpliva njegovo odločanje. Odločevalec se mora soočiti s številnimi zahtevami in jih tudi zadovoljiti. S preudarnim kompromisom poskuša združiti nasprotujoče si zahteve, tako da izoblikuje in združi interese, ki dano odločitev podpirajo. Odločitev je sprejeta, ko se nekaj ljudi, ki je vpletenih v proces odločanja, poenoti, da so našli rešitev. To naredijo z vzajemnim prilagajanjem in pogajanjem (Lesjak 1988).

Zaradi individualnih razlik med odločevalci in razlik v stabilnosti okolja se delež racionalnega proti intuitivnemu, proti preudarnemu načinu odločanja spreminja od enega odločevalca in odločitvene situacije do drugega. Združek načinov odločanja je torej odvisna od odločitvene situacije in odločevalca.

Večino odločitev sprejemamo tako, da se vsi trije načini odločanja med seboj prepletajo. Zato je verjetno najustreznejša zmes treh načinov odločanja, kar pripomore k boljšemu razumevanju odločanja in vedenja odločevalca. Aristotel je že davno zapisal, da smo ljudje zmes racionalnega in čustvenega. Prav tako vemo, da je okolje zmes sprememb in pritiskov, ki so po eni strani kaotični, po drugi pa jih je mogoče analizirati. Zato so takšne tudi odločitve, sprejete na tipično človeški način; pri tem pa se uporabljajo racionalne, zavestne analize in intuitivna, nezavedna vsebina. Učinkovit odločevalec združuje torej racionalen in čustven način odločanja, prav tako pa mora upoštevati tudi izvršljivost dane odločitve (Harrison 1975).

2.2.4 Odločitev in tveganje

Vsaka odločitev ima posledice, ki so lahko večje ali manjše. In prav zaradi posledic je lahko odločanje eno od najbolj stresnih opravil, ki jih moramo ljudje opraviti. Connor (2007) ugotavlja, da se večina ljudi odloča brez razmisleka o morebitnih posledicah, ki jih odločitev lahko povzroči, posledice pa so lahko zelo neprijetne (na primer izbira določene fakultete, ki se po nekaj zapravljenih letih in denarju izkaže kot napačna). Zato nekateri avtorji, na primer Birkova (2007), svetujejo, da bi morala biti odločitev, kako sprejemati odločitve, najpomembnejša v našem življenju. Priporoča, da se vsak posameznik, preden sprejme odločitev, vpraša:

- Kakšne pomisleke imamo: Kdo je nosilec odločitvenega procesa? Na koga odločitev vpliva? Smo že kdaj sprejemali podobne odločitve? Kakšne strategije smo uporabili? Ali so uporabljene strategije delovale?
- Kakšne posledice ima naša odločitev: Kakšen je najslabši mogoči scenarij naše odločitve? Kakšne posledice prinaša vsaka od različnih možnosti? Kakšne pridobitve prinaša vsaka od možnosti?

Poleg teh dveh vprašanj je treba vsake odločitvene razmere proučiti s finančnega, socialnega, pravnega, čustvenega, osebnega, družinskega, verskega in družbenega vidika.

Za uspešno odločanje potrebujemo prave podatke, to pomeni ustrezno količino kakovostnih podatkov. Napačno je razmišljanje, da bodo odločitve boljše, če bomo imeli več podatkov. Prevelika količina podatkov lahko odločevalca zmede, zato se ne odloči optimalno. Zagotoviti ustrezno količino kakovostnih podatkov za vse ravni odločanja je torej pomembna naloga vsake združbe. Prav nobena raven ne sme biti zapostavljena, tudi izvedbeni management ne. Po nedavni raziskavi TeraData (Terry 2006) 41 % izvedbenega managementa sprejema kritične poslovne odločitve in število pomembnih poslovnih odločitev vsakodnevno narašča. Pri odločitvah si lahko pomagamo tudi s »tradicionalnimi metodami«. Harris (1998) priporoča naslednje: T-graf, PMI (Plus-Minus-Interesting), Buridenovo metodo (metodo kritiziranja), metodo sodil (v nadaljevanju meril) z utežmi in odločitveno matriko.

Ljudje se vsak dan srečujemo s podatki, nekateri nam koristijo za takojšnje odločitve, drugi pa so morda pomembni za odločitve v prihodnosti. Enako je z združbami, ki se dobesedno utapljajo v podatkih. Kaj torej storiti s podatki, jih zavreči in s tem morda izgubiti koristno znanje, skrito v njih? Ti podatki bi morda lahko koristili združbi v prihodnosti pri pomembnih odločitvah in bi zmanjšali tveganje za negativne posledice. Pomembno je, da te podatke shranimo v zbirke podatkov, ki omogočajo kakovostno skladiščenje, hkrati pa hitro iskanje znanja, ko ga potrebujemo.

3 ZBIRKE PODATKOV KOT TEMELJ SHRANJEVANJA PODATKOV

Združba za poslovanje potrebuje kakovostne, pravočasne in zanesljive podatke o združbi (na primer podatki o zaposlenih, proizvodnji, prodaji, finančni podatki) in podatke o poslovnem okolju (na primer podatki o porabnikih, dobaviteljih, tekmecih). Podatke dobiva združba iz okolja (na primer podatki v okviru naročila izdelkov, reklamacije) ali v združbi (na primer delovni nalog, dobavnica). Treba jih je shraniti, in čeprav je relacijski model trenutno najbolj razširjen, ni edina možnost shranjevanja podatkov. Pred relacijskim modelom sta bila v uporabi dva modela, in sicer hierarhični ter mrežni podatkovni model, pojavljajo pa se tudi novi modeli, na primer objektni podatkovni model in hibridni modeli.

3.1 Hierarhični model

Hierarhične zbirke podatkov so bile prve med različnimi modeli, ki so bile množično razširjene. Predvsem so bile popularne v 60. in 70. letih dvajsetega stoletja, k čemur je največ pripomogel IBM s svojim sistemom za upravljanje informacij (angl. *Information Management System*). Temeljijo na načinu shranjevanja podobnih objektov za posamezen zapis. Objekti so shranjeni v obliki hierarhije, in sicer drevesne strukture (starš – potomec/angl. *parent – child*).

Večina hierarhičnih podatkovnih zbirk omogoča ponavljanje skupin zapisov (polja, ki lahko shranijo mnogo vrednosti). Značilnost hierarhičnih podatkovnih modelov je njihova hitrost in preprosta zasnova, pomanjkljivost pa se je zaradi načina shranjevanja (datotečni sistem – angl. *flat file*) pokazala v togosti. Največja omejitev po Kwanovem (2006) mnenju pa sta nezmožnost uvedbe relacij M : N (mnogo proti mnogo) in »šibka« referenčna integriteta. Avtorja Alex in Emil Vishnev (2005) pa navajata naslednje prednosti in slabosti hierarhičnih modelov:

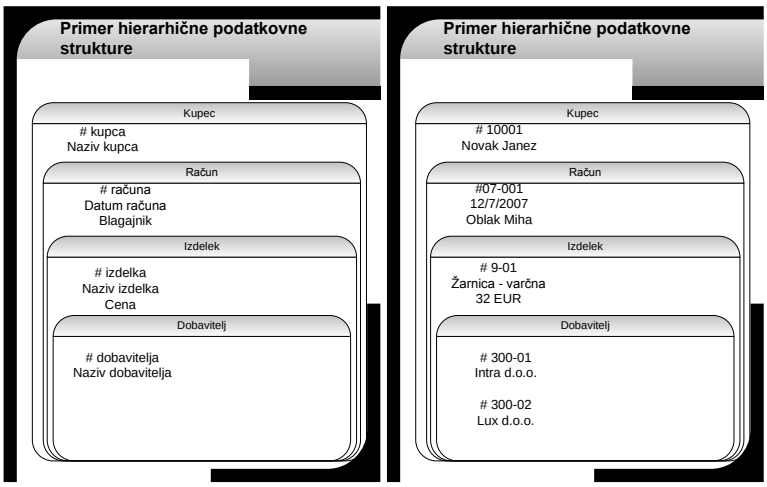
- prednosti: hitrost, podatkovna celovitost, sposobnost shranjevanja velike količine podatkov;
- slabosti: nujnost poznavanja fizične strukture, nezmožnost ad hoc poizvedb, odsotnost relacij M : N, majhne spremembe v strukturi podatkov so povzročale ogromno sprememb v poslovni logiki.

Na sliki 3 imamo primer hierarhične podatkovne strukture. Na primeru je zelo jasno videti, da relacije M : N ni mogoče postaviti (izdelek lahko dobavlja več dobaviteljev, medtem ko posamezen dobavitelj ne more dobavljati več izdelkov). Druga pomanjkljivost je »šibka referenčna integriteta«, ki jo v hierarhičnem modelu pravzaprav ni. S tem se dejansko misli stopnjo ponavljanja določenega zapisa. Lahko si predstavljamo, da ima kupec več naročil in

znotraj enega naročila več izdelkov. Za vsak zapis o izdelku bi morali navesti celotne podatke, tudi naziv in ceno. To pomeni, da bi bila stopnja ponavljanja določenega zapisa zelo visoka, in če bi nastala potreba po spreminjanju naziva določenega izdelka, bi imeli izredno zahtevno delo, da bi to tudi izvedli.

Navedene značilnosti jasno kažejo, da hierahičnemu modelu manjka kar nekaj lastnosti, ki bi pomagale k boljši kakovosti podatkov.

SLIKA 3: PRIMER HIERARHIČNE ZBIRKE PODATKOV



Vendar način hierarhičnega modela postaja spet popularen, in sicer s tehnologijo XML,[6] ki je izredno razširjena.

3.2 Mrežni model

Priljubljenost mrežnih zbirk podatkov sovpada s priljubljenostjo hierarhičnih zbirk podatkov. Potreba po relacijah M : N je pripeljala do mrežnega modela. Če je imel v hierarhičnem modelu potomec lahko le enega starša, je v mrežnem modelu lahko imel več staršev. Mrežni model je torej le nekakšna nadgradnja hierarhičnega modela, ki je bil uradno opredeljen leta 1971 na konferenci CODASYL (Olle 1978).

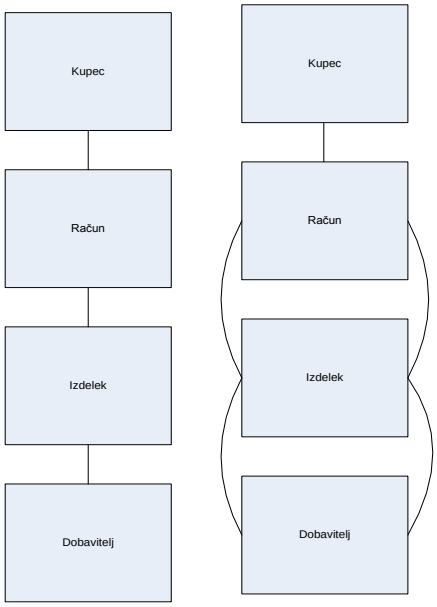
Glavni gradnik v mrežnem modelu je niz, ki ga sestavljajo ime, lastnik ter član zapisa. Relacija M : N je omogočena z možnostjo istega člana zapisa v več nizih. Za iskanje podatkov ni bilo treba treba »iti v globino«, zato je bilo iskanje izredno hitro, poizvedbe bolj zapletene. Vendar mrežna struktura prinaša tudi slabosti. Glede na to, da je mrežni model le nadgradnja hierarhičnega

[6] Sprva se je XML uporabljal kot sredstvo za izmenjavo podatkov med različnimi subjekti in objekti. Današnja uporaba pa je presegla začetne okvire in se uporablja tudi tam, kjer je ne bi pričakovali (na primer kot nastavitvena datoteka nekaterih Microsoftovih storitev).

modela, je od slednjega podedoval tudi večino njegovih slabosti, kot je potreba po poznavanju fizične strukture ter velik vpliv fizične izvedbe podatkovnega modela na uporabniško rešitev.

Vizualno je mrežni model enak kot hierarhični model, le da imajo lahko potomci več staršev (slika 4), na primer račun ima več izdelkov in več izdelkov je lahko na posameznem računu. Mrežnemu modelu prav tako manjkajo lastnosti, ki bi zagotavljale kakovostnejše podatke. Predvsem je problematično ponavljanje podatkov, ki zahteva natančno vzdrževanje in pušča več možnosti za napake.

SLIKA 4: HIERARHIČNI IN MREŽNI PODATKOVNI MODEL

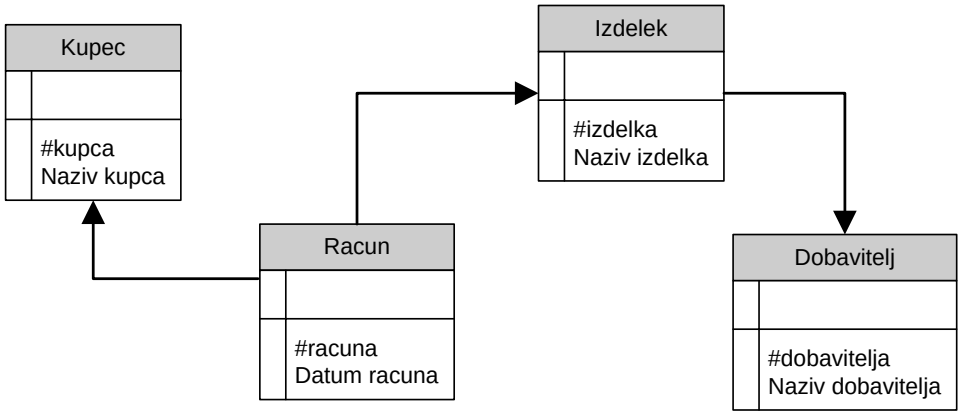


3.3 Relacijski model

Relacijski podatkovni model (slika 5) je nastal kot rešitev obstoječih težav hierarhičnega in mrežnega modela. Njegov utemeljitelj je Codd, ki je bil po poklicu matematik. Relacijski model temelji na teoriji množic. Središče relacijskega podatkovnega modela je preglednica, ki predstavlja entiteto – objekt, s stolpci, ki predstavljajo lastnosti/atribute entitet, in vrsticami, ki so n-terke. Vsaka preglednica ima enolično ime, ki jo ločuje od drugih preglednic, in v vsaki preglednici je stolpec, ki enolično določa posamezen zapis. Če želi uporabnik poiskati določen zapis, mora poznati samo ime preglednice in lastnost zapisa, ni mu treba poznati fizične strukture podatkovnega modela. Relacijski podatkovni modeli imajo tudi učinkovite SUBP (MS SQL, Oracle, DB2, Postgre, MySQL), ki še povečujejo njihovo priljubljenost. Za ravnanje s podatki v relacijskem podatkovnem modelu je bil razvit poseben jezik *Structured Query Language* (v nadaljevanju SQL), ki omogoča hitre operacije nad množicami podatkov, hkrati pa je izredno neučinkovit z vrstičnimi operacijami. Relacijski podatkovni model se je razvijal in se še razvija. Na

začetku ni poznal indeksov, ki jih lahko enačimo s kazalom v knjigi. S povečevanjem količine podatkov so operacije nad podatki postale časovno neučinkovite, zato je bilo treba poiskati rešitev. Pojavili sta se dve rešitvi, uvedba indeksov ter proces normalizacije/denormalizacije. Normalizacija je proces oblikovanja relacijske zbirke podatkov, kjer se preglednice oziroma relacije razdeljujejo na manjše enote, dokler posamezna relacija ne vsebuje lastnosti, ki so zelo povezane z glavnim ključem. Večina podatkovnih modelov je v tretji normalni formi (Ambler in Sadalage 2006).

SLIKA 5: PRIMER RELACIJSKE ZBIRKE PODATKOV



Pa si pobliže oglejmo prve tri normalne forme:

- prva normalna forma – relacija je v prvi normalni formi, kadar ne vsebuje ponavljajočih se lastnosti;
- druga normalna forma – relacija je v drugi normalni formi, kadar je v prvi normalni formi in kadar so vse lastnosti odvisne od celotnega ključa;
- tretja normalna forma – relacija je v tretji normalni formi, kadar je v drugi normalni formi in kadar med naključnimi lastnostmi ni nobenih odvisnosti.

Popolnoma normalizirana podatkovna zbirka je zelo prilagodljiva, vendar je časovna sledljivost podatkov v podatkovnem modelu težja, če ne celo nemogoča. Poleg tega je lahko učinkovitost popolno normaliziranega podatkovnega modela slaba (izbris, dodajanje, spreminjanje, prikazovanje podatkov). Denormalizacija predvsem zmanjša potrebo po različnih poizvedbah in združevanju preglednic. Kljub indeksom in normalizaciji pa se še vedno pojavljajo očitki o neučinkovitosti relacijske zbirke podatkov, še zlasti pri povezovanju številnih preglednic prek »dolgih« glavnih ključev. Drugi pogosto naveden očitek relacijskemu podatkovnemu modelu je odsotnost zapletenih podatkovnih struktur (polj). Če očitka premislimo, vidimo, da gre pri počasnosti pri povezovanju velikega števila preglednic verjetno bolj za slabšo fizično izvedbo – zaradi neznanja ali drugih vzrokov – kot pa za neučinkovitost modela samega. Poleg tega je

treba vedeti, da zmogljivosti računalniške opreme nezadržno rastejo, in kar je danes slabše učinkovito, bo jutri učinkovito. Tudi odsotnost zapletenih podatkovnih struktur ne more biti več slabost relacijskega podatkovnega modela, saj današnji SUBP-i omogočajo uvedbo izjemno zapletenih podatkovnih tipov. Zapletenih podatkovnih tipov na začetku res ni bilo v relacijskem modelu, vendar so nekateri avtorji, še posebno Christoper Date[7] (1995), že takrat poudarjali, da je problem izključno na izvedbeni ravni posameznih SUBP. Kljub vsemu napredku pa relacijski podatkovni model še zmeraj ne zadovoljuje vseh potreb na določenih področjih, zato je nastal objektni podatkovni model.

3.4 Objektni model

Objektni podatkovni model se je pojavil kot potreba po bolj zapletenih ureditvah v uporabniških rešitvah. Najprej se je uporabljal predvsem na področju CAD (*Computer Aided Development*) ter CAM (*Computer Aided Manufacturing*), danes pa se uporablja tudi na področju telekomunikacij, zdravstva, financ ter množičnih občil. Njegova velika prednost je v shranjevanju objektov in metod, ki nam omogočajo dostop do posameznega objekta (enkapsulacija). Glavna pomanjkljivost relacijskega podatkovnega modela je v »izkrivljanju« stvarnega sveta, ki jo objektni model odpravlja. To je najlaže ponazoriti s primerom zbirke podatkov o mačkah (slika 6). V objektnem podatkovnem modelu je laže slediti razvoju objekta skozi čas, kar je pomembna prednost v primerjavi z relacijskim podatkovnim modelom, kjer je ponazorjeno samo trenutno stanje. Prednost je tudi v dedovanju lastnosti posameznih objektov ter enotnem jeziku za programiranje in ravnanje z zbirko, kar pomeni, da je za isto funkcijo potrebno manj programiranja oziroma manj programske kode v primerjavi z relacijskim podatkovnim modelom. Zato je tudi vzdrževanje enostavnejše. S pojavom svetovnega spleta in spletnih uporabniških rešitev se je kot prednost izpostavila tudi navigacija, saj je premikanje z objekta na objekt bolj »naravno«. Cook (1997) v svojem članku ugotavlja, da je prednost tudi hitrost, kar pa Hand in Chandler (1998) izpodbijata.

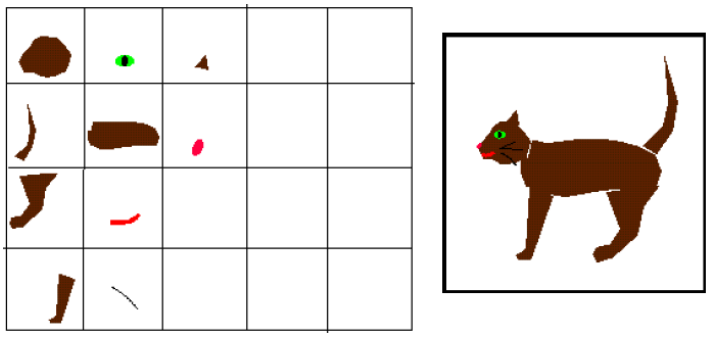
Cook (1997) namreč poudarja vzpostavitev instance posameznega objekta le enkrat, medtem ko vsaka poizvedba v relacijskem modelu ustvari svojo instanco. Hand in Chandler (1998) opozarjata na nevarnost »preglobokega« dedovanja, ki lahko upočasni dostop do posameznega objekta. Avtorja povzemata tudi druge slabosti objektnega podatkovnega modela:

- večja zapletenost – ni več ponazoritve s preglednicami;
- hitrost – hitrost dostopa do posameznih objektov je lahko zaradi dedovanja in združevanja zelo okrnjena;

[7] Christopher Date je bil ožji sodelavec Codda, očeta relacijskega modela, zato bi se njegove takratne trditve, ko je »branil« relacijski model, lahko zdele pristranske, vendar je danes – 10 let pozneje jasno, da se ni motil.

- pomanjkanje standardov – v relacijskem podatkovnem modelu je ANSII standard;
- pomanjkanje teoretične osnove – relacijski model temelji na teoriji množic.

SLIKA 6: PRIMER SHRANJEVANJA PODATKOV V RELACIJSKEM (LEVO) IN OBJEKTNEM MODELU (DESNO)



Vir: Hand in Chandler, Introduction to object-oriented databases, 1998.

3.5 Hibridni modeli

Zaradi reševanja zahtevnejših poslovnih in drugih problemov so nastajali različni modeli, ki so vsebovali prednosti posameznih podatkovnih modelov. V 80. letih prejšnjega stoletja se je pojavil relacijsko-mrežni model, ki je reševal probleme uporabniških rešitev CAD/CAM. O tem več govori Haynie (1981), ki poudarja vlogo »križancev« za reševanje zahtevnejših problemov. Združek relacijsko-mrežnega modela je nadomestil objektni model, tako da poglobljeno razmišljanje o relacijsko-mrežnem modelu ni smiselno.

Konec 90. let prejšnjega stoletja se je razvila razprava o relacijsko-objektnem podatkovnem modelu, ki pa je danes nekako zamrla. Zamisel je bila združiti prednosti relacijskega in objektnega podatkovnega modela. Objektni model naj bi »posodil« zapletenejše podatkovne tipe (časovne vrste, slika, video, avdio), enkapsulacijo (na primer shranjeni postopki), relacijski model pa naj bi prispeval razumljivost v smislu zgradbe (preglednice) ter robustnosti (transakcije). Ponudniki relacijskega SUBP so veliko funkcij objektnega modela vgradili v svoj sistem (shranjeni postopki, zapletenejši podatkovni tipi), zato je poimenovanje relacijsko-objektni podatkovni model nekako zamrlo.

Iz pregleda literature in izkušenj lahko trdimo, da je relacijski podatkovni model tisti, ki bo še nekaj časa osrednji podatkovni model za uporabniške rešitve. Opaziti pa je tudi, da ponudniki SUBP (Oracle, Microsoft, DB2 in drugi) spremljajo zahteve uporabnikov in svoje sisteme ves čas dopolnjujejo, predvsem z značilnostmi objektnega modela. Zato je še toliko bolj pomembno, da dobro poznamo možnosti, ki nam jih ponujajo omenjeni sistemi, in jih v praksi tudi uporabljamo. Z možnostmi je mišljena predvsem uporaba zapletenih podatkovnih tipov, ki omogočajo kakovostnejše FIRPM in posledično tudi kakovostnejše podatke za odločanje.

PREGLEDNICA 2: PREDNOSTI IN SLABOSTI POSAMEZNE PODATKOVNE STRUKTURE

	Hierarhični model	**Mrežni model**	**Relacijski model**	**Objektni model**	**Hibridni modeli**
Prednosti	Razumljivost Hitrost	Hitrost Zapletenejše poizvedbe	Zanesljivost Manjše podvajanje Večja prilagodljivost	Sledljivost Usmerjanje (v nadaljevanju navigacija) Zapletenejše ureditve	Razumljivost Prilagodljivost Hitrost Robustnost
Slabosti	Odsotnost relacij M : N Šibka referenčna integriteta Način shranjevanja	Šibka referenčna integriteta Način shranjevanja	Odsotnost zapletenih struktur Manjša učinkovitost	Hitrost Slabša razumljivost	

4 KAKOVOST PODATKOV

4.1 Pomembnost kakovosti in stroški nekakovostnih podatkov

Računalniško podprte zbirke podatkov imajo vedno večji vpliv na poslovanje organizacij (Xu 2003, 5), število napak v zbirkah podatkov pa je vedno večje (Klein 1998). Raziskave kažejo, da je v zbirkah podatkov približno 1–10 % nekakovostnih podatkov (Laudon 1986 ; Madnick in Wang 1992 ; Redman 1992), kar lahko resno ogrozi tekmovalno sposobnost združb. Novejše raziskave pa trdijo, da je v zbirkah podatkov sistemov sprotne obdelave podatkov delež nekakovostnih podatkov pogosto okoli 60–90 % (Dasu, Vesonder in Wright 2003), kar je izredno zaskrbljujoče, saj so zbirke podatkov izredno pomemben dejavnik za uspeh združb. Dokaz za to trditev je Olsonova (2003) ugotovitev, da so zbirke podatkov pravzaprav najpomembnejše premoženje združbe, vendar združbe tej vrsti premoženja ne namenjajo toliko pozornosti kot drugim vrstam premoženja. Olson (2003) prav tako ugotavlja, kako je management združb strpen do ogromnih količino nekakovostnih podatkov, ki po ocenah strokovnjakov povzročajo 15–20 % manjši dobiček v združbah.[8] Olsonu pritrjuje tudi Redman (2004), ki navaja, da nekakovostni podatki povzročajo 10 % manjši prihodek, vendar pozneje doda, da je ocena 20 % najbrž bolj pravilna. Odnos združba – zbirka podatkov je proučeval tudi Ambler (2006) v svoji raziskavi, v kateri ugotavlja, da sicer 96 % združb šteje zbirke podatkov za svoje premoženje, vendar le majhen delež združb z njimi dejansko ravna kot s premoženjem.

Kakovost podatkov je pomembna za vsako združbo, neodvisno od njene velikosti, zato bi morala biti skrb za kakovost podatkov ena od osrednjih nalog vsake združbe (Lee et al. 2002), vendar je skrb za kakovost navadno nizko na prednostni lestvici (Tayi in Ballou 1998). Podatki so še posebno pomembni v združbah, pri katerih podatke enačimo s poslovnimi učinki. Tako banke, zavarovalne ustanove in druge finančne institucije proizvajajo ogromne količine podatkov, ki so njihov osrednji poslovni učinek.

Podatki so temelj informacijskih sistemov in njihova natančnost je najpomembnejša razsežnost matrike kakovosti podatkov (Olson 2003, 3), o kateri je govora v nadaljevanju. Raziskava TDWI (Russom 2006) iz leta 2005 je pokazala, da imajo ameriške združbe vsako leto 600 milijard dolarjev stroškov zaradi nekakovostnih podatkov, pri čemer razsežnost natančnost podatkov zavzema največji delež v celotnih stroških. Raziskave so tudi pokazale, da se povprečna združba sicer zaveda problema nekakovostnih podatkov, vendar ta problem podcenjuje, saj nima

[8] Znano podjetje Dun & Breadstreet je moralo v 80. letih prejšnjega stoletja plačati 350.000 dolarjev odškodnine nekemu gradbenemu podjetju, ker je razglasilo, da je to podjetje bankrotiralo (Percy, 1986).

izdelanega stroškovnega modela in modela prihrankov, če bi ta problem odpravila (Friedman 2006).

TDWI je podobno raziskavo opravil že leta 2001 in takrat je 44 % anketiranih združb odgovorilo, da imajo težave z nekakovostnimi podatki. V letu 2005 je enako trdilo že 53 % združb, hkrati pa je delež združb, ki kakovosti podatkov niso posvečale pozornosti, v istem obdobju padel s 43 na 36 %. Raziskava v letih 2001 in 2005 je opozorila tudi na dejstvo, da problemi s kakovostjo podatkov vplivajo predvsem na informacijske projekte ter poslovne učinke združb. Veliko težav zaradi kakovosti podatkov je predvsem tehnične narave, na primer ureditev podatkov (85 %), zamude pri informacijskih projektih (52 %), težave s kakovostjo podatkov pa vplivajo tudi na poslovanje združb, nezadovoljstvo kupcev (69 %), sklepanje poslov (39 %), izgubo prihodkov (35 %), zvišanje stroškov (67 %) ter zmanjšanje ugleda oziroma verodostojnosti (77 %).

Nekakovostni podatki imajo velik vpliv na poslovanje združb. Burns (2005) ugotavlja, da nekakovostni podatki:

- povzročajo neposredne stroške: delo oseb, ki se ukvarjajo z odpravo nekakovostnih podatkov. Nekatere združbe imajo posebne oddelke, ki se ukvarjajo s strankami (na primer poslan napačen izdelek, izdelek poslan na napačen naslov);
- vplivajo na razmerja s strankami: isto stranko lahko imamo večkrat vneseno, zaradi česar večkrat prejme našo pošto; stranka ima lahko vpisan napačen naslov, zato ne dobi naročenega blaga. Vse to lahko povzroči izgubo stranke. Pri obravnavanju strank moramo upoštevati predvsem naslednja dejstva (Berson, Smith in Thearling 1999, 42):
    - Združbe porabijo veliko več denarja za pridobitev novega kupca kot za ohranitev že obstoječega kupca.
    - Veliko dražje je vnovič pridobiti kupca, ki je odšel k tekmecu, kot pa skrbeti, da ostane zadovoljen.
    - Izdelek je mnogo lažje prodati obstoječemu kot pa novemu kupcu.
    - Nekateri kupci združbi prinašajo več denarja kot drugi;[9]
- vplivajo na poslovne odločitve: slabi podatki so slaba osnova za sprejemanje izvedbenih, taktičnih in strateških odločitev;
- vplivajo na zamujene priložnosti: obvladovanje razmerij s kupci, poslovno obveščanje, kamor sodi SOZP, ne prinesejo želenega učinka. S SOZP lahko dobimo vzorce, ki ne ustrezajo dejanskemu stanju trga, združba pa zamudi svojo priložnost;
- negativno vplivajo na celotno kakovost poslovanja združbe;

[9] 5–15 % kupcev ustvarja večino dobička združbe, na drugi strani pa 25–46 % kupcev ustvarja le 1–5 % dobička, pri čemer zanje porabimo tretjino svojih prvin (Neal 2001).

- so življenjsko pomembni: v bolnišnicah je napačen zapis krvne skupine pacienta lahko usoden, posledica pa je smrt.

Olson (2003) navaja še dva pomembna vpliva nekakovostnih podatkov, in sicer:

- ponovitev napak z uvedbo novih sistemov – izdelava skladišč podatkov in drugih sistemov, ki jemljejo podatke iz obstoječega sistema, povzroča, da se nekakovostni podatki »naselijo« tudi v novih sistemih oziroma veliko projektov niti ne uspe. V raziskavi *Meta Group* iz leta 1999 je bilo ugotovljeno, da 41 % projektov izdelave skladišč podatkov ne uspe zaradi nekakovostnih podatkov (Oliveira, Rodrigues in Henriques 2006).
- zmanjšana proizvodnja zaradi težav v oskrbni verigi (ang. *supply chain management*) – zaradi nekakovostnih podatkov se v proizvodnjo v določeni oskrbni verigi lahko dostavi napačen proizvod ali pa se proizvod dostavi z zamudo. Takšne težave navadno povzročajo stroške v proizvodnji, saj se slednja lahko celo zaustavi.

Vse navedene posledice nekakovostnih podatkov povzročajo izgubo časa in energije ljudi, ki se ukvarjajo z odpravo teh posledic, vendar sčasoma to postane del vsakdanjih opravil, kar vpliva na mnenje, da je to nekaj normalnega. Posledično ti problemi navadno niti ne dosežejo strateškega managementa v združbi, ki bi se nanje verjetno odzvali oziroma bi se morali odzvati. Redman (1998) vplive nekakovostnih podatkov na povprečno združbo razdeli glede na raven managementa, in sicer:

- izvedbeni management – slabše zadovoljstvo strank, višji stroški, nezadovoljstvo zaposlenih. Stranke pogosto ne odpuščajo »preprostih« napak (na primer dostava napačnega poslovnega učinka).[10] Stroški lahko dosegajo 10 % prihodkov, v storitvenih združbah pa tudi do 50 %. Odpravljanje napak je delo, ki zaposlenih ne zadovoljuje, zato postanejo nezadovoljni;
- taktični management – sprejemanje slabših odločitev, daljši proces odločanja, nezaupanje v združbo;
- strateški management – težave pri načrtovanju strategije, odvrnitev pozornosti od pomembnih opravil.

Stroške nekakovostnih podatkov zelo podrobno opredelijo tudi English (1999), Loshin (2001) ter Eppler in Helfert (2004). Opredelitev je prikazana v preglednici 3.

[10] Podjetje Business Link London (BLL) je s čiščenjem naslovov v svoji zbirki podatkov zmanjšalo podvajanje naslovov strank na samo 2,5 %, kar je pomenilo 100.000 funtov prihranka. Veliko stroškov so imeli, ker so pošiljali trženjsko gradivo ljudem, ki sploh niso obstajali (McCue 2006).

PREGLEDNICA 3: OPREDELITEV STROŠKOV NEKAKOVOSTNIH PODATKOV

English	Loshin	Eppler in Helfert
Stroški zaradi neizvedbe procesa Nepovratni stroški Stroški izgube ugleda Stroški ukvarjanja z nezadovoljnimi strankami **Stroški popravljanja napak** Ukvarjanje s podvojenimi podatki Iskanje manjkajočih podatkov Manjša produktivnost Stroški ponovnega opravljanja dela Stroški preverjanja podatkov Stroški reprogramiranja programske opreme Stroški čiščenja in popravljanja podatkov **Izgubljene in zgrešene poslovne priložnosti** Stroški izgubljenih poslov Stroški zamujenih poslovnih priložnosti Stroški zmanjšanja vrednosti združbe zaradi predhodnih dejstev	**Stroški operativnih procesov** Stroški zaznavanja napak Stroški popravljanja napak Stroški ponovnega opravljanja dela Stroški preprečevanja napak Stroški uveljavljanja jamstva Zmanjšanje prihodkov **Stroški taktičnih in strateških procesov** Stroški zamud Stroški predkupne pravice Stroški nedela Stroški zgrešenih priložnosti Stroški zaradi izgube zaupanja v združbo Stroški vzdrževanja	**Neposredni stroški** Stroški preverjanja podatkov Stroški vnovičnega vnašanja podatkov Stroški izravnav (kompenzacij) **Posredni stroški** Stroški izgube ugleda Stroški zaradi napačnih odločitev Stroški izgubljenih vlaganj

Vir: prirejeno po Batini in Scannapieca, Data quality: concepts, methodologies and techniques, 2006.

Englisheva (1999) opredelitev stroškov nekakovostnih podatkov ima tri glavne kategorije:

- stroški neizvedbe procesa so stroški, ki nastanejo, ko nekakovostni podatki povzročijo, da se posamezen proces ne izvede pravilno, na primer nenatančni podatki o naslovih povzročijo, da se proces komuniciranja ne more izvesti pravilno;
- stroški popravljanja napak so stroški, ki jih zahteva ravnanje s procesom odprave napak, na primer nezaupanje v podatke povzroča vnovično preverjanje zaposlenih, ki ne verjamejo v podatke;
- stroški izgubljenih in zgrešenih poslovnih priložnosti pa se nanašajo na stroške, ki nastanejo, ker združba zaradi nekakovostnih podatkov ni izvedla določenih vlaganj ali pa je že dobljene posle izgubila, na primer trženjska akcija zaradi nekakovostnih podatkov o naslovih strank ne more doseči že obstoječih strank, kar posledično pomeni manjše prihodke.

Loshinova (2001) razlaga je pravzaprav podrobnejša opredelitev Redmanove razdelitve stroškov nekakovostnih podatkov z večjim poudarkom na taktični in strateški ravni, medtem ko Eppler in Helfert (2004) uporabljata metodo »od spodaj navzgor«. Njuno izhodišče so stroški, ki jih navaja

literatura, te stroške pa sta potem razdelila v dve skupini, in sicer med neposredne ter posredne stroške nekakovostnih podatkov.

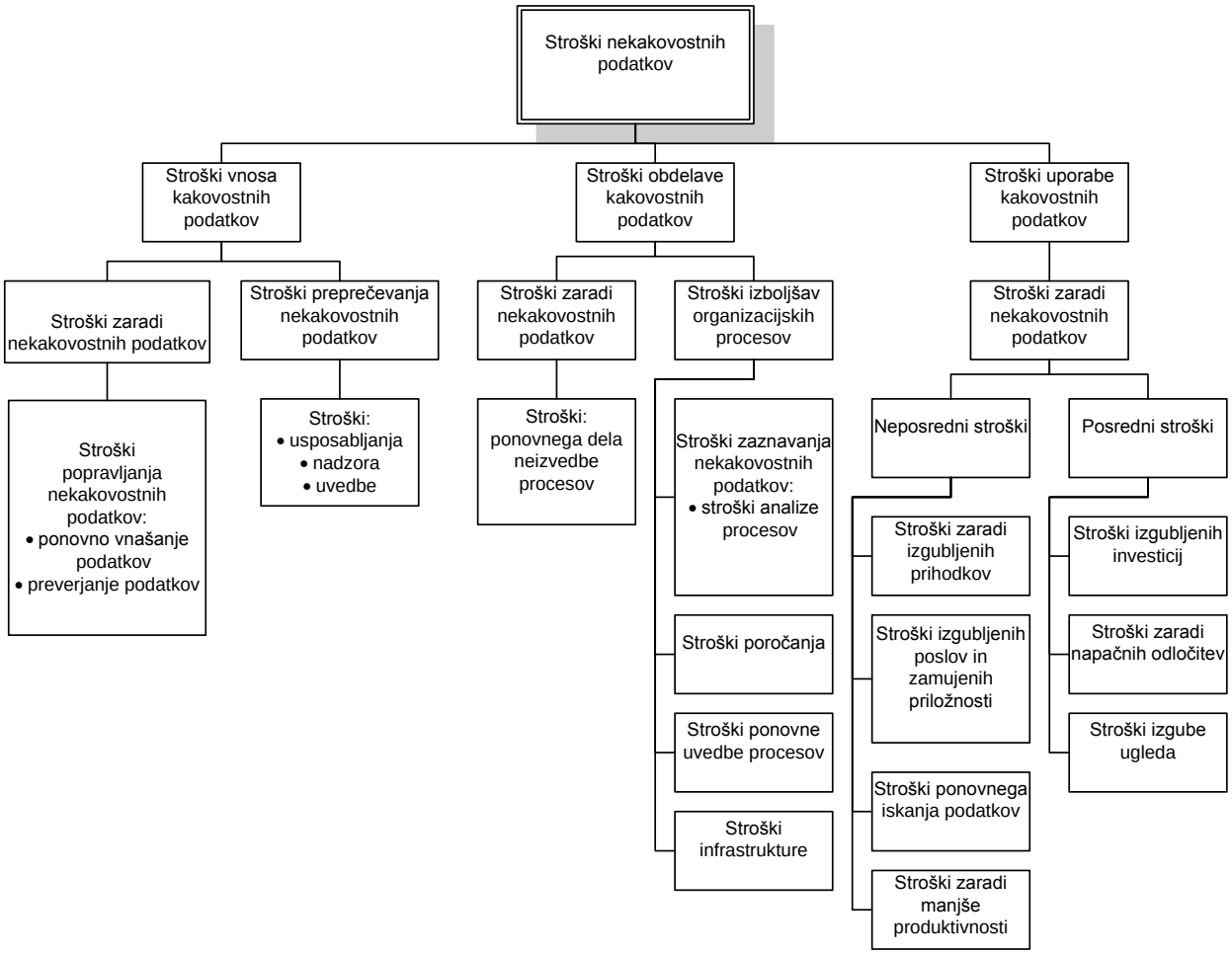
Kakovost podatkov ni nekaj samoumevnega in brezplačna, ampak je treba zanjo nekaj postoriti. Cinična, a hkrati popolnoma umestna je pripomba Karasconyja (2006), da je najboljši način za zagotavljanje kakovosti podatkov proizvajanje kakovostnih podatkov. Eppler in Helfret (2004) ugotavljata, da gre pri zagotavljanju in izboljšanju kakovosti podatkov predvsem za stroške, ki nastanejo zaradi želje po preprečevanju nekakovostnih podatkov, za stroške odkrivanja nekakovostnih podatkov ter za stroške popravljanja, medtem ko se Loshin (2001) in English (1999) usmerjata predvsem na stroške za izboljšanje procesov. Batini in Scannapieca (2006) v svojem delu *Data Quality* povzemata razdelitve vseh štirih prej omenjenih avtorjev in jih združita v enotno porazdelitev stroškov, pri čemer njuna razdelitev upošteva tako stroške kot posledico nekakovostnih podatkov kot tudi stroške za zagotavljanje oziroma izboljšanje kakovosti podatkov. Njuno razdelitev prikazuje slika 7.

Združbe se za odpravo nekakovostnih podatkov oziroma izboljšavo obstoječe kakovosti podatkov odločajo iz ekonomskih razlogov. Pri določanju ukrepov je treba pretehtati stroške, ki jih bodo ti ukrepi zahtevali, pa tudi koristi, ki jih bodo ukrepi prenesli. Ekonomski izračun ne sme biti vodilo, ko se odločamo o zagotavljanju kakovosti podatkov širšega družbenega in moralnega pomena (na primer zdravstveni podatki, podatki za zagotavljanje nacionalne varnosti). Že prej omenjena avtorja English (1999) in Loshin (2001) razvrstita koristi kakovostnih podatkov v tri skupine:

- koristi, ki so neposredno merljive v denarju (na primer boljša kakovost podatkov vpliva na višje prejemke). Te koristi lahko poimenujemo tudi denarne koristi (angl. *monetizable*);
- merljive koristi (angl. *quantifiable*), ki jih sicer ne moremo neposredno izraziti v denarju, lahko pa jih merimo v neki drugi številski merski enoti (na primer boljši podatki davčne uprave omogočajo časovni prihranek združb in posameznikov, ki ga lahko izmerimo. Med merljive koristi spadajo časovno skrajšanje poslovnega ciklusa, večja produktivnost ter večji tržni delež;
- neotipljive koristi (angl. *intangible*) so koristi, ki jih ne moremo neposredno izraziti z nobeno številsko mersko enoto (na primer izguba ugleda združbe). Neotipljive koristi izboljšanja kakovosti podatkov so večje zadovoljstvo strank, večje zadovoljstvo zaposlenih ter boljša kakovost storitev.

Oba avtorja, English (1999) in Loshin (2001), se strinjata glede denarno ovrednotenih koristi, medtem ko je pri neotipljivih koristih Loshin bolj skromen, English pa bolj obširen, pri čemer poudarja predvsem pomembnost boljše kakovosti storitev.

SLIKA 7: RAZDELITEV STROŠKOV NEKAKOVOSTNIH PODATKOV



Vir: *prirejeno po Batini in Scannapieca, Data quality: concepts, methodologies and techniques, 2006.*

V raziskavi TDWI v letu 2005 (podrobneje predstavljena na str. 4-5), so raziskovali tudi koristi kakovostnih podatkov (Russom 2006). Raziskava je pokazala, da kakovostni podatki povečajo zaupanje v analitične sisteme[11] (76 %), zmanjšajo potrebo po različnih obdelavah (70 %), poenotijo opredelitve entitet – ena resnica (69 %), povečajo zadovoljstvo strank (57 %), znižajo stroške (56 %) ter povečajo prihodke (30%). Poleg navedenih pozitivnih učinkov je raziskava obravnavala tudi vpliv na donosnost naložb in proračun združb. Kar 43 % anketiranih združb je odgovorilo, da vlaganje v kakovost podatkov pozitivno vpliva na donosnost vlaganj, medtem ko je pozitiven učinek kakovosti podatkov na proračun združb prepoznalo kar 80 % združb. Od navedenih 80 % pritrdilnih odgovorov je 42 % anketirancev odgovorilo, da bo proračun ostal enak, 31 % jih je menilo, da se bo povečal zmerno, medtem ko jih je 7 % menilo, da se bodo prihodki zaradi višje kakovosti podatkov močno povečali.

[11] V oklepajih so navedeni deleži anketiranih združb v ZDA, ki so pritrdile posamezni koristi.

Pregledali smo torej stroške nekakovostnih podatkov, stroške zagotavljanja kakovosti podatkov ter koristi kakovostnih podatkov za združbo. Iz tega lahko sklepamo, da imajo podatki velik vpliv na poslovanje združb, zato jim je treba posvetiti veliko pozornosti. Zakaj podatki imajo tako velik vpliv, pa bomo pogledali v naslednjem podpoglavju.

4.2 Kakovost podatkov in različni informacijski sistemi

Vidik kakovosti podatkov se spreminja z vrsto informacijskih sistemov oziroma namenom uporabe podatkov, zato je v naslednjih podpoglavjih obravnavan vidik SSOP, skladišča podatkov, SOZP ter obvladovanja razmerij s kupci.

4.2.1 Kakovost podatkov in sistemi sprotne obdelave podatkov

Ko govorimo o SSOP, imamo v mislih celovite poslovne informacijske sisteme. Te Markus in Tanis (2000, 176) opredelita kot sisteme, ki omogočajo povezavo med transakcijskimi podatki in poslovnimi procesi v združbi. Glavni namen SSOP je podpora in povezovanje poslovnih procesov ter obdelava podatkov v realnem času za vso združbo (Strong in Volkoff 2004).

Orr (1998) opredeli kakovost podatkov v informacijskih sistemih kot kakovost preslikave podatkov realnega sveta v informacijski sistem, natančneje v zbirko podatkov. V svojem članku zavrača tezo o 100-odstotni kakovosti podatkov in hkrati dodaja, da se bo tudi v primeru 100-odstotne kakovosti podatkov v določenem trenutku kakovost v odvisnosti od časa zmanjševala. Zato je pomembno, da združba zagotavlja takšno kakovost podatkov v SSOP, ki ji omogoča sprejemanje odločitev za preživetje združbe (govori o razsežnostih natančnost in pravočasnost). Orr (1998) tudi poudarja spreminjanje realnega sveta kot največjo težavo pri zagotavljanju kakovosti, saj se realni svet hitro spreminja, podatki v SSOP pa so statični.

Na tem mestu se postavi vprašanje odgovornosti za ažuriranje podatkov. Po Orrovi raziskavi se odgovornosti otepajo tako razvijalci informacijskih sistemov kot tudi uporabniki. Prvi izražajo stališče, da je njihova naloga zagotoviti ustrezen informacijski sistem v smislu doslednega podatkovnega modela in uvedbe poslovnih pravil v uporabniško rešitev, medtem ko skrb za podatke prelagajo na uporabnike. Ti pa poudarjajo svoje nepoznavanje informacijskega sistema kot posledico njegove zapletenosti in kot glavne krivce navajajo razvijalce. Resnica je nekje vmes. Razvijalci morajo zagotoviti kakovostno preslikavo stvarnega sveta, vključno s poslovnimi pravili, uporabniki pa morajo biti vestni pri svojem delu in morajo slediti organizacijskim predpisom glede ravnanja s podatki, če ti predpisi obstajajo.

Pri kakovosti podatkov je dobro upoštevati nekaj pravil (Orr 1998):

- Podatkom, ki jih ne uporabljamo, se kakovost poslabša.

- Kakovost podatkov v informacijskih sistemih je funkcija njihove uporabe, torej moramo kakovost proučevati z vidika namena uporabe.
- Kakovost podatkov se poslabša s starostjo informacijskega sistema.
- Če je neka lastnost v podatkovnem modelu dokaj trajna in ni pričakovati, da se bo spremenila, ima njena dejanska sprememba veliko negativnih posledic v informacijskem sistemu.
- Vse navedene lastnosti veljajo tudi za meta podatke (podatki o podatkih).

Na žalost pa se navedena pravila redko upoštevajo in združbe zbirajo podatke »na zalogo«, torej z možnostjo prihodnje uporabe, čeprav se to redkokdaj zgodi. Zato je priporočljivo zbirati le tiste podatke, ki jih potrebujemo zdaj ali pa jih bomo potrebovali v bližnji prihodnosti. Z uporabo je mišljena dejanska uporaba podatka, in ne le da so podatki v »vlogi statista« pri določenem poročilu ali na ekranski sliki. Pogosto se namreč zgodi, da se določeni podatki samo navidezno uporabljajo, dejanska uporaba oziroma uporabna vrednost pa je zanemarljiva.

Za izboljšanje kakovosti podatkov v informacijskih sistemih uporabljamo različne metode, ki so predstavljene v nadaljevanju, vendar metode ne pomagajo zadosti, če ne spremenimo poslovnih procesov. Hall (2005) svetuje oblikovanje oddelčnih oziroma organizacijskih predpisov, ki se ukvarjajo izključno s kakovostjo podatkov. To pomeni porazdelitev odgovornosti za ravnanje s podatki med vnašalce, končne uporabnike, revizorje in druge, kar zahteva spremembe v razmišljanju ljudi.

4.2.2 Kakovost podatkov in skladišča podatkov

Razvoj informacijskih sistemov je največ pozornosti namenil SSOP in transakcijskim podatkom. S povečevanjem obsega poslovanja združb se je v SSOP zbiralo preveč podatkov, ki so vplivali na manjšo hitrost delovanja informacijskih sistemov. Posledično se je pojavila misel o skladišču podatkov, kamor bi shranili podatke, ki jih ne uporabljamo vsak dan oziroma katerih vrednost se ne spreminja pogosto. Ti podatki so analitični podatki in jih uporabljamo za različne analize poslovanja združbe. Pomembna razlika med obema informacijskima sistemoma je v cilju, ki ga ta dva sistema zasledujeta. SSOP je pomembna predvsem učinkovitost delovanja (hitrost, celovitost), analitičnemu sistemu pa prilagodljivost. Pri načrtovanju skladišča podatkov je zlasti pomembna izdelava podatkovnega modela skladišča podatkov. Izdelava podatkovnega modela skladišča podatkov mora potekati neodvisno od obstoječega podatkovnega modela v SSOP. Podatkovni model skladišča bo verjetno manj normaliziran kot podatkovni model SSOP, poleg tega pa bo slednji verjetno vseboval več ponavljajočih se oblik določenega podatka (informacija o določenem poslovnem učinku bo verjetno v SSOP uporabljena v različnih oblikah, skladišče podatkov pa mora pojavne oblike združiti).

Eden od ciljev skladišča podatkov je doseči, da so analize čim bolj prilagojene različnim profilom uporabnika. Skladišče podatkov naj bi bilo dostopno čim večjemu številu končnih uporabnikov, seveda pa ne moremo pričakovati, da bo s skladiščem podatkov ustrezno ravnal vsak končni uporabnik. Manj tehnično usposobljeni uporabniki bodo morda zadovoljni že z uporabo razpredelnice in preprostih poizvedb, tehnično bolj usposobljeni uporabniki pa se bodo zadovoljili le z najmočnejšimi orodji za analizo. Med te spadata OLAP in SOZP. Slednjemu je namenjeno naslednje podpoglavje, zato se bomo tukaj osredotočili samo na OLAP. Rešitve OLAP nam omogočajo hitre analize velikih količin podatkov, ki so večrazsežnostno organizirani in po navadi obsegajo daljše časovno obdobje. Podatke navadno črpamo iz skladišč podatkov, kamor pritekajo iz različnih virov, tako notranjih kot tudi zunanjih. Čeprav je tehnologija OLAP največkrat omenjena kot orodje za management srednje ravni (taktični management), jo uporabljamo tudi na ostalih dveh ravneh kot pripomoček pri odločanju, zato je kakovost podatkov v njej izredno pomembna. Prvi pogoj za kakovostne odločitve s pomočjo tehnologije OLAP so kakovostni vhodni podatki, torej podatki v skladišču podatkov, kamor jih prenesemo iz SSOP. Pri prenosu moramo biti izredno dosledni, hkrati pa je to po navadi trenutek resnice o stanju naših podatkov v SSOP in pogosto bomo zelo začudeni, kako nekakovostne podatke imamo, predvsem v smislu natančnosti in popolnosti. Vse navedene pomanjkljivosti imajo velik vpliv na analize, ki temeljijo na njih. Manjkajoč podatek o območju stranke nam onemogoča analizo na primer prodaje po območjih, vendar nas moderni sistemi OLAP že opozarjajo na takšne napake, zato je treba večjo pozornost posvetiti napačnim vrednostim. Tako se lahko naša največja stranka preseli v drugo območje, mi pa podatka ne posodobimo, zato bi bila analiza časovnih vrst za prizadeto območje napačna.

Ferengul (2006) opozarja predvsem na tri vrste posledic nekakovostnih podatkov v skladiščih podatkov, in sicer na slabšo učinkovitost poslovanja, zmanjšanje donosnosti vlaganj ter pravne posledice. Posebno občutljive so po njegovem mnenju pravne posledice, saj se skladišča podatkov vedno pogosteje uporabljajo za finančna poročila, ki vplivajo na odločanje managementa.

Iz navedenega sledi, da je treba pri uvajanju skladišča podatkov in analiz, ki izhajajo iz njih, največjo pozornost nameniti pripravi in čiščenju podatkov. Po različnih raziskavah (Manning 1999 ; Neely 2005) ukvarjanje s podatki pomeni od 60 do 80 % vseh stroškov uvedbe skladišča podatkov. To je v skladu tudi z mojimi izkušnjami in »nekaj« klikov v uporabniški rešitvi za zgraditev sistema za poslovno obveščanje je manjši del potrebnega truda v celotnem projektu. Na to, kako pomembna je kakovost podatkov za uspeh projekta o skladišču podatkov, opozarja tudi Hudicka (2003).

Kakovost podatkov je mogoče z različnimi programskimi rešitvami sicer izboljšati, vendar je to po navadi le kratkotrajna rešitev. Najpomembnejša je kakovost podatkovnih virov, tj. navadno SSOP, in boljši je podatkovni model SSOP, manj težav je pri prenosu podatkov, kar spet potrjujejo moje lastne izkušnje. Vendar so kljub prizadevanju za izboljšanje kakovosti podatkov v skladiščih podatkov ta po Ferengulu (2006) še vedno izpostavljena petim različnim dejavnikom tveganja:

- preslaba kontrola kakovosti podatkov v vseh korakih gradnje skladišča podatkov – združbe po navadi izvedejo kontrolo le na enem mestu, kar posledično pomeni popravljanje slabih podatkov (kurativa) namesto odkrivanje vzroka nekakovostnih podatkov (preventiva);
- nezadostna kontrola finančnih in drugih številčnih podatkov. Podatke v skladiščih podatkov moramo na različne načine preverjati z izhodiščnimi podatki. Predvsem moramo biti pozorni pri različnih pretvorbah (merske enote, valute), saj so tovrstni podatki pogosto tarča posrednega preverjanja zakonodajalcev;
- odsotnost oziroma pomanjkljiva revizija postopkov tokov informacij v združbi;
- podcenjevanje staranja podatkov – počasi, a vendarle spreminjajoče/starajoče se razsežnosti podatkov, ki zahtevajo popravljanje podatkov, shranjenih v skladiščih podatkov;
- pomanjkljivi celostni zunanji mehanizmi kontrole kakovosti podatkov – združbe navadno storijo napako, ker so vzpostavljeni mehanizmi kontrole kakovosti podatkov del projekta skladišča podatkov namesto del celotnega poslovanja združbe. Zaradi tega je veliko kontrolnih mehanizmov ciljno omejenih.

Poleg analitičnih rešitev OLAP iz skladišča podatkov črpamo tudi podatke za SOZP.

4.2.3 Kakovost podatkov in odkrivanje zakonitosti v podatkih

SOZP je pojem, ki ga je v zadnjih letih sprejela širša množica ljudi. Z njegovo priljubljenostjo se je spreminjala tudi opredelitev, vendar še zmeraj lahko rečemo, da je odkrivanje zakonitosti v podatkih obsežen skupek tehnologij za podporo odločanja, od preprostih poizvedb do umetne inteligence, z namenom odkriti vzorce, povezave, spremembe, anomalije, medsebojne odvisnosti v podatkih, shranjenih v primerni obliki, da bi pridobili znanje, ki pomaga ljudem pri odločanju.

SOZP ima svoje korenine v statistiki, ki je postavila temelje večini tehnologij, na katerih temelji odkrivanje zakonitosti v podatkih. Statistika vsebuje zasnutke, kot so standardni odklon, standardna porazdelitev in varianca, regresijska analiza, analiza skupine, intervali zaupanja. Te zasnutke uporabljamo pri analizi podatkov in povezav med podatki.

Naslednji korak pri razvoju SOZP je bil umetna inteligenca (angl. *artificial intelligence*), ki pomeni hevristično metodo reševanja problemov. Ker ta metoda zahteva veliko procesorsko moč, se je začela razvijati šele v začetku 80. let prejšnjega stoletja. Programska oprema za to metodo je bila tako draga, da je bila dostopna le večjim znanstvenim zavodom in državnim uradom.

Zadnji korak v razvoju SOZP pa pomeni metoda, ki se imenuje strojno učenje (angl. *machine learning*). Ta metoda pomeni povezave med statistiko in umetno inteligenco, programska oprema je bila veliko cenejša kot pri umetni inteligenci, poleg tega tudi bolj napredna. Ko so programi obdelovali podatke, so se zraven tudi učili, tako da so lahko dali različne napotke za odločanje, pač glede na kakovost obdelanih podatkov.

V čem se statistika in SOZP pravzaprav razlikujeta? Odgovor morda na prvi pogled ni očiten, saj statistiko danes uporabljamo kot tehniko SOZP za določanje tržnih segmentov in predvidevanje vedenja kupcev. Treba pa je tudi dodati, da je dolgo časa veljalo, da je SOZP del statistike in je bila beseda »data mining« označena kot slabšalnica (Pregibon 1997, 7).

Pomembna razlika med statistiko in SOZP je v njuni ciljni skupini. Statistika ni bila nikoli mišljena kot orodje končnega uporabnika, SOZP pa je namenjen prav slednjemu. Orodja SOZP avtomatizirajo statistične procese, poleg tega pa so bolj stabilna in namenjena večjim količinam podatkov. Če je za uporabo statistike potreben statistični analitik, ki postavi različne hipoteze in jih potem preverja, se pri SOZP vzorci oziroma povezave odkrijejo samodejno. Zato lahko rečemo, da statistiko obvladuje človek (angl. *human driven*), SOZP pa obvladujejo podatki (Grossman et al. 1998).

Zakaj torej odkrivati vzorce med podatki? Odgovor je preprost: zaradi koristi, ki nam jih omenjeni proces prinaša, in zaradi količine podatkov, ki jih imajo združbe v sedanjosti na voljo. Če bi imeli majhno število podatkov (na primer nekaj deset strank), potem bi lahko že s preprostim pregledom ugotovili značilnosti podatkov. Toda če je teh podatkov ogromno (GB in TB), potem je takšen način popolnoma neprimeren oziroma nemogoč. Zato so se pojavila orodja, ki nam omogočajo odkrivanje vzorcev v podatkih na ogromnem številu podatkov. Proces odkrivanja vzorcev v podatkih pogojujejo predhodne izkušnje in poznavanje podatkov, količina in kakovost podatkov.

SOZP vsebuje različne algoritme, ki jih Witten in Frank (2000) razdelita na enostavne in zapletene. Med enostavne algoritme uvrščata statistiko, razvrščanje ter metodo najbližjega soseda, med zapletenejše algoritme pa prištevata odločitvena drevesa, indukcijo pravil, nevronske mreže ter genetske algoritme.

Vendar niso vsi problemi primerni za reševanje s SOZP. Kennedyjeva (1997) opredeljuje potrebne značilnosti, ki pomagajo razjasniti dilemo, ali problem reševati s SOZP ali ne. Po njenem mnenju so te značilnosti naslednje (Kennedy 1997):

- velika količina podatkov, ki se navezujejo na obravnavani problem;
- problem ni jasen oziroma niso jasne njegove značilnosti;
- problem se lahko opredeli kot vhodni/izhodni problem;
- reševanje s tradicionalnimi matematično-statističnimi metodami temelji na preveč predpostavkah oziroma je nepopolno.

Velikokrat je za reševanje problemov priporočljiv tudi združek s tradicionalnimi matematično-statističnimi metodami, kar še poveča možnost boljše dobljene rešitve, priporočljivo pa je razrešiti tudi nekaj praktičnih vprašanj:

- Kako dober mora biti rezultat oziroma kakšna odstopanja bomo še dopuščali?
- S katerimi obstoječimi rešitvami problema bomo dobljeni rezultat primerjali?
- Katere podatke bomo uporabili za preverjanje uporabljenih modelov?

Pri reševanju problemov je dobro začeti z manjšimi problemi, da si pridobimo izkušnje, nato pa preidemo k reševanju zapletenejših problemov.

Podatke, ki jih potrebujemo za reševanje problemov s SOZP, dobimo iz skladišča podatkov. Vendar ti »surovi« podatki po navadi niso primerni za takojšnjo obdelavo, temveč jih moramo pregledati in odpraviti morebitne napake. Največkrat je treba poiskati izjemne vrednosti, manjkajoče, neusklajene in nepopolne vrednosti podatkov. Izjemne, manjkajoče, neusklajene ter nepopolne vrednosti podatkov so pogost pojav v skladiščih podatkov in količina takšnih podatkov lahko zasede kar veliko prostora v skladišču podatkov. Nepopolni podatki se lahko pojavijo iz več razlogov, in sicer lahko manjkajo lastnosti podatkov, ki nas zanimajo, določeni podatki niso bili vključeni v skladišče podatkov zaradi slabe analize problema ipd. Izjemne vrednosti so največkrat posledica slabih virov podatkov, ki take vrednosti pošiljajo podatkovnemu skladišču, včasih pa so izjemne vrednosti iskana rešitev, na primer iskanje poneverb. Naštete anomalije odpravimo v procesih čiščenja, povezovanja, spreminjanja in izbrisa podatkov.

Več anomalij (slabša kakovost podatkov) imamo v podatkih, več truda je treba vložiti v obdelavo podatkov. In nasprotno, manj anomalij imamo (boljša kakovost podatkov), manj truda je treba vložiti v obdelavo podatkov. Kakovost podatkov je zelo pomembna za uspešen proces odkrivanja vzorcev med podatki.

4.2.4 Kakovost podatkov in obvladovanje razmerij s kupci

Podatki imajo temeljno vlogo tudi v procesu, imenovanem obvladovanje razmerij s kupci (angl. *Customer Relationship Management – CRM*), ki je svojo največjo priljubljenost dosegel na prelomu tisočletja, zdaj pa je opaziti malo manjše navdušenje, čeprav je zdaj na trgu množica uporabniških rešitev, ki naj bi podpirale oziroma omogočale obvladovanje razmerij s kupci.

Kaj obvladovanje razmerij s kupci je? Obvladovanje razmerij s kupci je vse, kar je povezano z iskanjem, pridobivanjem in ohranjanjem kupcev. Lahko torej rečemo, da je obvladovanje razmerij s kupci poslovna strategija, ki v sredino poslovanja postavlja kupca in mora zato temu prilagoditi razmišljanje zaposlenih, poslovne procese ter tehnologijo (2007). Pomembno je, da zaznamo potrebe kupca, še preden jih slednji izrazi. Rud (2001) navaja naslednje cilje obvladovanja razmerij s kupci:

- pridobitev novih kupcev – obvladovanje razmerij s kupci nam omogoča učinkovitejše trženjske akcije, ki bodo pritegnile več novih kupcev;
- pridobitev dobičkonosnih kupcev – novi pridobljeni kupci naj bi imeli visoko življenjsko vrednost;
- izogibanje tveganim kupcem – za združbo je seveda visok strošek, če pridobi novega kupca, ta pa ne plača blaga oziroma storitev. Obvladovanje razmerij s kupci nam z izdelavo različnih modelov omogoča, da se takšnim kupcem izognemo;
- povečanje dobičkonosnosti že obstoječih kupcev;
- zadržanje dobičkonosnih kupcev;
- vnovična pridobitev kupcev – kupce, ki so nekoč že bili zvesti naši združbi, a so pozneje odšli k tekmecu, lahko spet pridobimo nazaj;
- povečanje zadovoljstva kupcev;
- povečanje prodaje;
- znižanje stroškov – bolj učinkovito trženje.

Bistvo pri obvladvanju razmerij s kupci je kupec, ki mu združba prilagodi razmišljanje svojih zaposlenih, tehnologijo ter poslovne procese. Ta proces je pravzaprav že zelo star. Pa si poglejmo zakaj? Pred velikim razmahom potrošniške družbe so ljudje po navadi ves čas hodili kupovat blago oziroma storitve k istemu trgovcu. Slednji je kmalu poznal vsakega posameznika, predvsem v smislu, kaj posamezen kupec rad kupi, katere izdelke po navadi kupi skupaj ipd. Ko sta se trgovec in kupec bolje spoznala, je trgovec že lahko predvidel kupčeve potrebe ter temu prilagodil svojo ponudbo. Toda z rastjo trga in prehodom v potrošniško družbo se je stik med trgovcem in kupcem pretrgal. Kmalu je bil posamezni kupec le še anonimen kupec, ki je kupoval v kakšni veliki trgovski verigi. Za trgovca je bilo razumljivo, da so se njegovi poslovni učinki

prodajali, veljalo je prepričanje, da če bo kakšen kupec šel k tekmecu, bo morda tudi kakšen kupec, ki zdaj hodi k tekmecu, prešel k njemu. Toda z razvojem trga je takšno razmišljanje kmalu vodilo v propad. Počasi so začeli trgovci spoznavati, da je izjemno pomembno, da skrbijo za svojega kupca, ga poskušajo spoznati, predvsem njegove navade pri nakupovanju, ter ohraniti porabnika zadovoljnega, da ne bo odšel k tekmecu. Jim Novo (2004) govori o dveh profilih, in sicer o demografskem ter o vedenjskem. Demografski profil opisuje kupca, podaja opisna dejstva, vedenjski profil pa nam pove, kaj kupec počne. Treba se je vprašati, kateri profil se nam zdi pomembnejši. Pravzaprav sta pomembna oba profila, pomembnost pa je odvisna od razmer, v katerih smo. Če bi radi prodali oglasni prostor ali pripravili ustrezno vsebino spletne strani, je prvi profil pomembnejši. Drugi profil pa je povezan z vedenjskimi navadami in je pomembnejši v razmerah, ko nas zanima, kaj naš kupec počne. Na tem mestu bi rad poudaril, da je dolgoročno najbrž pomembnejši vedenjski profil kupca, saj iz tega lažje sklepamo na vedenje posameznega kupca v prihodnje. Kupca, ki je prenehal obiskovati našo stran in se ne bo več vrnil, prav nič več ne zanima, ali je na primer spletna stran prilagojena njegovim potrebam. Potrebno je torej, da kupca ne izgubimo.

Obvladovanje razmerij s kupci poteka prek osmih korakov (Postma 1999):

- izbira kupcev: najprej ocenimo vrednost kupca, in sicer izračunamo pričakovano povprečno vrednost življenjske dobe kupčeve zvestobe. Za združbo je najpomembnejše, da pridobi in obdrži najdonosnejše kupce, ki niso nujno največji, ampak so lahko srednje veliki. Dobičkonosni kupec je tista oseba, gospodinjstvo ali združba, ki prinese dohodek, ki presega stroške, ki jih ima združba, da pridobi, proda in postreže temu kupcu (Kotler 1994, 52). Veliki kupci po navadi zaradi velikih količin dobijo tudi velik popust in s tem se njihova dobičkonosnost zmanjša. Na drugi strani pa majhni kupci zaradi majhnih količin sicer plačajo poln znesek, a so izvedbeni stroški poslovanja z njimi tako visoki, da se dobičkonosnost spet zmanjša. Podatki, ki so pomembni za izračunavanje dobičkonosnosti kupca, so lahko notranji in zunanji, pri čemer vedno najprej uporabimo notranje podatke. Moramo pa biti pazljivi pri podatkih v smislu pravilnosti. Za ocenjevanje posameznega kupca lahko uporabimo še zunanje podatke, na primer o trgu, kjer ta kupec nastopa, kako obvladujemo trg;
- prilagoditev ponudbe posameznemu kupcu: pri tem je pomembno, da smo predhodno oblikovali profil kupca s pomočjo podatkov iz trženjske zbirke podatkov. Če teh podatkov ni na voljo, če torej ne moremo oblikovati profila posameznega kupca, ga poskušamo uvrstiti v določeno skupino z enakimi značilnostmi;
- izbira komunikacijske poti oziroma občila: v praksi je najboljše, da najprej presodimo funkcije (poosebljenje, interaktivnost, draženje čutil) in stroške posameznega

komunikacijskega občila, velikokrat pa se združbe odločajo, da bodo najprej poskusile z najcenejšimi in postopoma prehajale na dražje komunikacijske poti;

- izvedba in zaznavanje: pomembno je, da celotno akcijo dobro načrtujemo in po sami izvedbi tudi spremljamo potek. Pri tem lahko uporabimo vrsto trženjskih raziskav, ki pa lahko dajo nenatančne podatke;
- odgovarjanje je odziv na našo trženjsko akcijo, torej ko kupci naročijo naše blago oziroma storitev ali pa samo dajo neko informacijo (na primer kupon: da/ne). Vrsta odgovora je odvisna od medija, ki smo ga uporabili za našo trženjsko akcijo. Če smo kupcem poslali pošto, v kateri je bil kupon, bo odgovor različen, kot če s kupci komuniciramo prek interneta;
- izpolnitev storitve oziroma posredovanje blaga, ki smo ga tržili v akciji. Paziti je treba, da izpolnimo obljube, ki smo jih navedli v določenemu občilu, kot so dobavni roki, kakovost, cena, zaloga.
- Povratna informacija o izkušnji je zelo pomembna za povečanje obsega znanja o kupcih, ki nam v prihodnosti lahko zelo pomaga pri obvladovanju razmerij s kupci. Informacije, ki jih dobimo v procesu obvladovanja razmerij s kupci, zapišemo v trženjsko zbirko podatkov ter z različnimi orodji za analizo izdelamo natančnejši profil posameznega kupca. Tako imamo možnost, da bo prihodnjič naša trženjska akcija uspešnejša, saj bo znanje o kupcu ter primernih občilih za uporabo večje.

Iz navedenih korakov sledi, da so kakovostni podatki izjemno pomembni v vsakem koraku obvladovanja razmerij s kupci, kar v svojem delu trdi tudi Postma (1999), ki pravi, da je obvladovanje razmerij s kupci trženjski proces, sestavljen iz tehnologije in podatkov, ki jih Berson, Smith in Thearling (1999, 360) razdelijo v tri skupine:

- podatki, ki povedo, kdo pravzaprav naš kupec je, so *opisni podatki* (starost, spol, naslov, dohodek na člana gospodinjstva) in se ne spreminjajo pogosto;
- podatki, ki opisujejo trženjske akcije, ki smo jih opravili in so bile usmerjene k potencialnim in obstoječim kupcem. Ti podatki se navadno nanašajo na tip komunikacije, občilo, ki je bilo uporabljeno za komunikacijo, datum komunikacije, stroške komunikacije;
- podatki o odzivu kupca, torej *transakcijski podatki*. Ta vrsta podatkov se spreminja zelo pogosto.

To pomeni, da bi se morale združbe zelo dobro zavedati stanja svojih podatkov, preden se odločijo za uvedbo obvladovanja razmerij s kupci. Z njim povezani projekti vse prevečkrat propadejo zaradi odsotnosti primerne vizije in strategije, ki ne vključuje zagotavljanja kakovosti podatkov v celotnem procesu oziroma slabšajo učinkovitost obvladovanja razmerij s kupci.

PREGLEDNICA 4: OBVLADOVANJE RAZMERIJ S KUPCI

Obvladovanje razmerij s kupci		
Podatki Zbirke podatkov Odkrivanje zakonitosti v podatkih Kartica kupca	Mediji: Internet Klicni centri Avtomatski odzivniki Interaktivna televizija	Trženjski proces: 1 na 1 trženje Interaktivno trženje E-trgovina

Vir: prirejeno po Postma, The new marketing era: marketing to the imagination in a technology-driven world, 1999.

Dravis (2003) navaja primer združbe, ki je z zvišanjem kakovosti podatkov v sistemu obvladovanja razmerij s kupci na leto prihranila 54.000 dolarjev. Hay (2003) opozarja na težave, na katere lahko naletimo pri uvedbi obvladovanja razmerij s kupci:

- Podatki o strankah so pogosto shranjeni v zbirkah podatkov, ki ne poznajo referenčne integritete ali pa ta ni bila uvedena.
- Odsotnost referenčne integritete lahko povzroča slabšo kakovost podatkov.
- Težave pogosto nastopijo tudi pri opredelitvi entitete stranka.
- Podatki v različnih sistemih so pogosto različno podrobni. V določenem sistemu imamo lahko dnevne podatke, v drugem tedenske, v tretjem mesečne, ki jih je treba pretvoriti na skupni imenovalec.
- Polja oziroma lastnosti entitet imajo v različnih sistemih sicer lahko enaka ali podobna imena, vendar je njihova poslovna vsebina popolnoma drugačna (v določenem polju se za označevanje uporabljata 1 in 0, v drugem sistemu pa na primer D in N). Spet je treba biti pozoren pri povezovanju takšnih podatkov, hkrati pa se v navedenem primeru pokaže nepogrešljivost kakovostnega meta podatkovnega modela.
- Težave nastopijo tudi, če polja sploh ne vsebujejo podatkov.

Na koncu je treba ostro nasprotovati mnenju, da je obvladovanje razmerij s kupci poslovni učinek, ki ga je mogoče kupiti in uvesti v poslovanje posamezne združbe. Vse prevečkrat je mogoče videti osebe, odgovorne za trženje, ki obvladovanje razmerij s kupci enačijo s klicnimi centri, kar je seveda popolnoma zgrešeno. Obvladovanje razmerij s kupci je filozofija, proces, ki ga je treba neprestano izboljševati in ki za svoje delovanje potrebuje kakovostne podatke.

4.3 Razsežnosti kakovosti podatkov – matrika kakovosti podatkov

Kakovost podatkov je zapleten in obsežen pojem, katerega reševanje je po mnenju Dasuja in Johnsona (2003) izrazito interdisciplinarno. Pravita, da je zagotavljanje kakovosti podatkov

sestavljeno iz oblikovanja organizacijskih predpisov različnih strokovnjakov z mnogo področij. Vendar pa je tudi pri oblikovanju predpisov treba upoštevati Paretovo načelo, saj z nekaj predpisi lahko zagotovimo okoli 50-odstotno kakovost vsak naslednji predpis pa ima vpliv na manjši delež podatkov. Avtorja v svojem delu *Exploratory Data Mining and Data Cleansing* kritizirata splošno razširjene opredelitev kakovosti podatkov zaradi statičnosti, ki jo te opredelitve ponujajo, vendar je treba poudariti, da se pojavljajo tudi bolj dinamične opredelitve.

Izraz kakovost podatkov je osrednja tema raziskovanja na treh področjih (Wang, Storey in Firth 1995) na področju kakovosti podatkov, na področju ugotavljanja uspešnosti informacijskih sistemov in njihovih uporabnikov ter na področju računovodstva in revizije.

Kaj je torej kakovost podatkov? V literaturi se pojavljajo številne opredelitve kakovosti podatkov. Tako avtorja Anne Marie Smith (2006) in Sid Adelman (2006) opredeljujeta kakovost podatkov kot:

- delež veljavnosti – kolikokrat je vrednost določenega elementa v naboru veljavnih vrednosti;
- delež popolnosti – kolikokrat je vrednost podatka sploh navedena;
- delež natančnosti – kolikokrat se pojavi vrednost podatka, ki je točna in poslovno uporabna;
- delež stanovitnosti – primerjava vrednosti podatkov skozi čas.

Kljub različnim opredelitvam izraza kakovost podatkov ne moremo prezreti dejstva, da so pionirsko delo na tem področju opravili Morey (1982), Ballou in Pazer (1985), Delone in McLean (1992) ter Wang in Strongova (1996). Morey (1982) je raziskoval kakovost informacijskih sistemov in ugotavljal vzroke za napake v njih, opredelil pa je tudi razsežnost natančnost. Večina napak po njegovem mnenju izhaja iz napak v postopkih ter zakasnelih popravkih. Ballou in Pazer (1985) sta opredelila štiri razsežnosti, in sicer: natančnost, popolnost, doslednost ter pravočasnost, ki jih je med prvimi uporabil Laudon (1986) za razlago problemov podatkov o kriminaliteti v ZDA. Delone in McLean sta leta 1992 opredelila naslednje najpomembnejše razsežnosti kakovosti podatkov: natančnost, pravočasnost, zanesljivost, popolnost in ustreznost. Njuno opredelitev so dolgo časa povzemali številni avtorji (Rudra in Yeo 1999), v današnjem času pa se bolj nagibajo k izrazu kakovost podatkov, kot sta ga opredelila Wang in Strongova.

Wang in Strongova (1996) sta se lotila raziskovanja razsežnosti kakovosti podatkov z vidika uporabnikov. V svoji raziskavi sta leta 1996 postavila temelje za vse nadaljnje opredelitve, ki večinoma nadgrajujejo njune ugotovitve. Z anketiranjem končnih uporabnikov sta opredelila 20 skupin, ki sta jih pozneje združila v štiri skupine, v katere sta uvrstila ugotovljene razsežnosti kakovosti podatkov. Za merjenje kakovosti podatkov sta opredelila naslednje razsežnosti:

natančnost, objektivnost, zaupanje, ugled, razlagalnost, razumljivost, jedrnatost, doslednost, koristnost, pravočasnost, popolnost, količina, dostopnost in varnost. Navedene razsežnosti kakovosti podatkov sta združila v štiri skupine: osrednje razsežnosti, predstavljajoče razsežnosti, kontekstualne razsežnosti in razsežnosti dostopnosti.

PREGLEDNICA 5: RAZSEŽNOSTI KAKOVOSTI PODATKOV

Skupina	Razsežnost
Osrednje razsežnosti	Natančnost, objektivnost, zaupanje, ugled
Predstavljajoče razsežnosti	Razlagalnost, razumljivost, jedrnatost, doslednost
Kontekstualne razsežnosti	Ustreznost, pravočasnost, popolnost, količina
Razsežnosti dostopnosti	Dostopnost, varnost

Vir: Wang in Strong, Beyond accuracy: what data quality means to data consumers, 1996.

Eden od avtorjev, ki nadgrajuje delo Wanga in Strongove, je tudi Olson (2003). Pri njegovi opredelitvi kakovosti podatkov je pomemben vidik nameravane uporabe. To pomeni, da na podatke ne gleda samo statično, temveč vedno uporabi vidik nameravane uporabe. Kot je že bilo omenjeno, se kakovost podatkov proučuje tudi z vidika informacijskih sistemov ter zadovoljstva končnih uporabnikov. Halloran in soavtorji (1978) v svojem članku govorijo, kako razvoj PIS vpliva na njegovo kakovost, in opredelijo razsežnosti uporabnost, zanesljivost in neodvisnost. Bailey in Pearson (1983) sta zasnovala matriko za ugotavljanje zadovoljstva uporabnikov z informacijskim sistemom. Opredelila sta 39 različnih dejavnikov, med katerimi so tudi natančnost, zanesljivost, pravočasnost in popolnost. Zadnje področje, ki poskuša opredeliti kakovost podatkov, je področje, ki je neposredno najmanj povezano s kakovostjo podatkov, tj. področje računovodstva in revizije. Računovodstvo in revizija kot področji zahtevata visoko stopnjo zanesljivosti in natančnosti podatkov, pri čemer sta z natančnostjo mišljeni velikost in pogostost napak. Feltham (1968) dodaja še pravočasnost in ustreznost kot zaželeni lastnosti podatkov na področju računovodstva in revizije.

V literaturi je torej mogoče zaslediti veliko število razsežnosti kakovosti podatkov, vendar se številni raziskovalci ne morejo poenotiti pri opredelitvi najpomembnejših razsežnosti, tj. katera razsežnost je bolj in katera manj pomembna, ter pri opredelitvi posamezne razsežnosti (niti razsežnost natančnost ni enotno opredeljena). Za opredelitev razsežnosti so predlagane različne možnosti:

- opredelitev razsežnosti na osrednje in zunanje (Wang in Strong 1996),
- opredelitev razsežnosti s stališča teorije informacij (Delone in McLean 1992).
- opredelitev razsežnosti s stališča trženja in iz tega izhajajočega kupca (Wang in Strong 1996).

- opredelitev razsežnosti s stališča uporabnika in nameravane uporabe (Olson 2003).

V svoji knjigi sem za opredelitev kakovosti podatkov uporabil matriko kakovosti podatkov (najpomembnejše razsežnosti) po Wangu in Strongovi (1996), ki jo bom nadgradil z Olsonovo (2003) metodo nameravane uporabe (kjer je to smiselno in mogoče), predstavil pa bom tudi metodo Dasuja in Johnsona (2003), ki zavračata dosedanje opredelitve izraza kakovost podatkov in ponujata procesno opredelitev.

Opredelitve izraza kakovost podatkov so se lotile tudi različne institucije, na primer kanadski statistični urad, ki navaja naslednje razsežnosti matrike kakovosti podatkov: ustreznost, natančnost, pravočasnost, dostopnost, razlagalnost in razumljivost (Quality 2006).

4.3.1 Natančnost

Natančnost je opredeljena kot razdalja med vrednostma »v« in »v'«, pri čemer »v'« pomeni pravo vrednost iz realnega sveta, ki jo »v« poskuša prikazati (če je osebi ime Janez, potem je vrednost »v'« = Janez in je pravilna, medtem ko je »v« = Jnez in posledično nepravilna), Ballou (Ballou in Pazer 1982, 1985, 1987) pa opredeljuje natančnost s podobnostjo zapisanega podatka in dejanske vrednosti.

Po mnenju Batinija in Scannapieca (2006, 20) obstajata dve vrsti natančnosti, in sicer skladenska (v nadaljevanju sintaktična) ter pomenska (v nadaljevanju semantična) natančnost.

S sintaktično natančnostjo merimo odnos med vrednostjo »v« in vrednostmi množice elementov, ki jo »v'« lahko pripišemo. V tem primeru bi bila vrednost »v« = Miha pravilna, četudi je »v'« = Janez, kajti vrednost »Miha« je v množici vrednosti, ki jo »v'« lahko zasede. Razdaljo med »v« in elementi merimo z različnimi primerjalnimi funkcijami (na primer stopnja popravljanja, ki vsebuje najmanjše število potrebnih brisanj in vstavljanj znakov, potrebnih za spremembo zapisa »s« v »s'«).

V preglednici 6 imamo naštete nekatere modele proizvajalca Honda, od katerih je podatek Cvic sintaktično nepravilen, saj njegove vrednosti ni v naboru mogočih vrednosti (Jazz, Civic, Accord, S2000 …). Najbliže temu podatku je vrednost Civic in za pretvorbo vrednosti Cvic v Civic je potrebno vstavljanje črke i, torej je potreben 1 korak popravljanja.

Drugi vidik natančnosti je semantična natančnost, ki pa je težje določljiva kot sintaktična natančnost. Pri semantični natančnosti ugotavljamo razdaljo med vrednostjo »v« in pravo oziroma iskano vrednostjo »v'«. Če v navedeni preglednici pogledamo prvi zapis, in sicer lastnost proizvajalec, vidimo, da je proizvajalec modela Jazz Toyota, kar je sintaktično natančno (Toyota

je v množici avtomobilskih proizvajalcev), je pa semantično nenatančno, saj je pravi proizvajalec modela Jazz Honda.

Semantično natančnost je najlažje opisovati z DA/NE ali PRAVILNO/NEPRAVILNO. Da bi lahko določili semantično natančnost, moramo poznati pravo vrednost, ki jo iščemo, ali pa moramo biti vsaj sposobni potrditi ali zavreči trditev, da je »v« pravilna vrednost. Pri določanju semantične natančnosti nam lahko pomaga tudi vedenje o vzrokih sintaktične nenatančnosti. Če slednja izhaja večinoma iz tipkarskih napak, potem z doseganjem sintaktične natančnosti navadno dosegamo tudi semantično natančnost. V nasprotnem primeru pomeni zagotoviti semantično natančnost iskanje istih podatkov v različnih podatkovnih virih ter iskanje pravilnega podatka. Slednje pa je povezano s problemom iskanja oziroma opredelitve istih entitet. Pri takšnem načinu zagotavljanja semantične natančnosti se moramo odločiti, ali dve n-terki pripadata isti entiteti oziroma objektu realnega sveta.

PREGLEDNICA 6: PRIMER SINTAKTIČNE NATANČNOSTI IN NENATANČNOSTI

Model avtomobila proizvajalca	Proizvajalec	Sintaktična natančnost	Št. korakov popravljanja
Jazz	**Toyota**	Da	0
Cvic	**Honda**	Ne	1
Accord	**Honda**	Da	0

Poleg natančnosti posamezne lastnosti (kot v primeru, navedenem zgoraj) določamo tudi natančnost posamezne povezave (relacije) ter celotne zbirke podatkov. Ko pojmovanje natančnosti razširimo s posamezne lastnosti (atributa), se pojavijo nova vprašanja, na primer vprašanje podvajanja (ista n-terka je vnesena večkrat), ki pa jih je z dobrim podatkovnim modelom mogoče razrešiti. Na tem mestu se spet potrjuje predpostavka o vplivu podatkovnega modela na kakovost podatkov.

Pri opredeljevanju natančnosti se po navadi izračuna razmerje med pravilnimi podatki ter vsemi podatki (Lee 2006 ; Redman 2005), vendar pa je popolnoma jasno, da omenjena opredelitev ni zadostna. Te nezadostnosti se lotevajo Fisher, Lauria in Matheus (2007). V svojem delu navajajo očitno razliko, kako so napake porazdeljene. Več kot očitno namreč je, da je potrebno več napora za odpravo napak, ki so enakomerno porazdeljene v na primer celotni preglednici, kot za odpravo napak, ki so samo v eni vrstici ali samo v enem stolpcu, čeprav je izračunano razmerje enako v vseh treh navedenih primerih.

Na določanje semantične natančnosti posamezne n-terke imajo velik vpliv lastnosti te n-terke. Te lastnosti so lahko »usodne« in preprečujejo iskanje istih vrednosti v drugih virih podatkov (predpostavljamo, da je napačna davčna številka, ki enolično določa zavezanca), če je njihova

vrednost nepravilna, ali pa imajo le manjši vpliv na iskanje istih n-terk (na primer lastnost starost določene n-terke ima nepravilno vrednost).

Razsežnost »natančnost« je treba proučiti tudi z vidika nameravane uporabe. Pri tem izhajamo s problemskega področja, ki so mu namenjeni obravnavani podatki. Kako bi torej določili natančnost podatkov z vidika nameravane uporabe? Najlažje je to pokazati s konkretnim primerom. Recimo, da imamo zbirko podatkov prebivalcev določenega kraja, v kateri pa je mnogo napak (napačna imena, starost …). V tem hipotetičnem primeru predpostavljamo, da je približno 20 % podatkov napačnih oziroma nenatančnih. Torej, ali so ti podatki z vidika natančnosti kakovostni ali ne? Če želimo omenjeno zbirko podatkov uporabiti za pošiljanje vabil na občinske volitve, potem je zbirka podatkov nekakovostna oziroma nenatančna; v primeru nameravane uporabe za trženjsko akcijo določene združbe pa je podatkovna zbirka x v smislu natančnosti dovolj kakovostna.

Podatke v smislu natančnosti lahko izboljšamo z dobrim podatkovnim modelom in ustreznim načrtovanjem uporabniške rešitve. Spet se osredotočimo na primer vnosa podatkov o modelih in proizvajalcih avtomobilov, ki smo ga predhodno obravnavali. V uporabniški rešitvi bi lahko namesto navadnega vnosnega polja uporabili spustni seznam in tako zagotovili sintaktični vidik natančnosti podatkov. Poleg tega bi glavni spustni seznam (proizvajalci) lahko povezali s spustnim seznamom modelov ter tako zmanjšali možnost za semantično nenatančnost. Vendar to ne preprečuje napak pri neposrednem vnosu v podatkovni model. V tem primeru bi morali uporabiti zapletene lastne podatkovne tipe, ki bi vsebovali podobna poslovna pravila.

Kljub enostavnosti in razumljivosti razsežnosti natančnost ostaja kar nekaj odprtih vprašanj, pri katerih se raziskovalci ne morejo poenotiti:

- Kaj je dejanska vrednost (vrednost, ki jo zapisan podatek poskuša predstavljati)?
- Ali morajo biti vsa polja, ki sestavljajo zapis, natančna oziroma ali morajo odsevati dejansko vrednost, da zapis opredelimo kot natančen?
- Ali naj se zapis, ki ima eno polje nenatančno, opredeljuje kot bolj natančen v primerjavi z zapisom, ki ima dve polji nenatančni?

4.3.2 Doslednost

Doslednost kot razsežnost kakovosti podatkov pomeni enak način predstavljanja podatkov skozi čas in prostor, kar omogoča usklajenost in primerljivost podatkov v različnih časovnih obdobjih in iz različnih virov (Bertin 2004 ; Chapman 2005 ; Redman 1996). Batini in Scannapieca (2006) opredeljujeta doslednost kot usklajenost podatkov s semantičnimi pravili, opredeljenimi v relacijskih preglednicah in datotečnih sistemih. Awad in Gotterer (1992) pa namesto doslednosti

opredelita nedoslednost, in sicer pravita, da je nedoslednost pojavljanje različnih inačic istih podatkov v zbirki podatkov.

Redman (1996) opredeljuje dve vrsti doslednosti: semantično doslednost ter strukturno doslednost. Semantična doslednost preprosto pomeni, da so podatki vedno v istih poljih, zato je njihovo iskanje poenostavljeno. Zagotovitev semantične doslednosti je po Chapmanu (2005) mogoča z ustreznim FIRPM, čemur bi bilo treba dodati vlogo meta podatkovnega modela z ustreznimi opisi in opredelitvami posameznih polj. Strukturna doslednost pa opredeljuje pojavljanje/zapis podatka vedno v enakem formatu. V preglednici 7 imamo primer semantične in strukturne nedoslednosti. Naziv gospod/gospa bi namreč moral biti vedno v polju Naziv (semantična doslednost) in vedno v enaki obliki/formatu (gosp., ga. je nepravilno – strukturna doslednost).

PREGLEDNICA 7: PRIKAZ SEMANTIČNE IN STRUKTURNE NEDOSLEDNOSTI

Naziv	Priimek
Gospod	Novak
	Gospod Meglič
Gosp.	Turk
Gospa	Antončič
Ga.	Mihelič

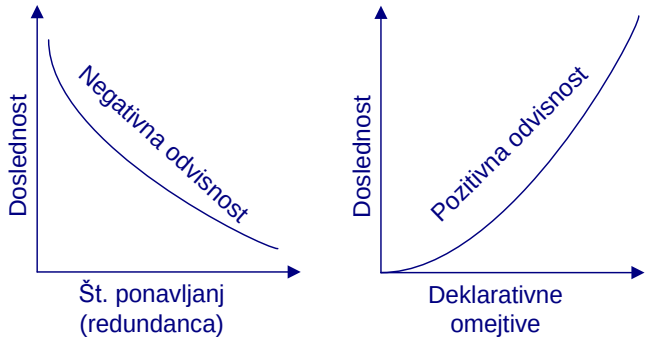
V skladu z Batinijevo in Scannapiecovo (2006) opredelitvijo doslednosti in ob upoštevanju opredelitve relacijskega podatkovnega modela lahko govorimo o deklarativnih omejitvah kot o sredstvu za zagotavljanje doslednosti, ki je razsežnost kakovosti podatkov. Poglejmo si primer: preglednica »Zaposleni« vsebuje podatke o nazivu, delovni dobi in osebnem dohodku. Poslovno pravilo pravi, da ima zaposleni z več kot tremi leti delovne dobe mesečni osebni dohodek 1500 evrov, poleg tega pa imamo v združbi določen tudi najvišji mogoči osebni dohodek, in sicer 3000 evrov. Ti dve poslovni pravili bi lahko uveljavili kot deklarativno omejitev nad poljem osebni dohodek in tako zagotovili semantično doslednost. Poleg pravkar navedenih pravil pa obstajajo tudi logična pravila, na primer osebni dohodek ne more biti negativen oziroma nižji od zajamčenega osebnega dohodka (v Sloveniji država predpisuje zajamčeni osebni dohodek, ki ga država jamči zaposlenemu).

Deklarativne omejitve so obširno proučevali in so sestavni del vsakega sodobnega SUBP (Scannapieca, Missier in Batini 2005). Pomembno je, da takšne deklarativne omejitve uvedemo na posameznih poljih ali skupinah polj in sporočila prenesemo končnemu uporabniku prek uporabniškega vmesnika. Deklarativne omejitve na poljih omogočajo visoko raven semantične

doslednosti tudi v primeru neposrednega dostopa do podatkov v zbirki podatkov in s tem pripomorejo k večji kakovosti podatkov.

Poleg pravil v relacijskem podatkovnem modelu lahko semantična pravila uveljavimo tudi v drugih okoljih, na primer na področju statistike. In takšna pravila tudi obstajajo in jih precej uporabljajo, na primer pri opisnih vprašanjih, kadar sta podatka zakonski stan in starost vsebinsko povezana, ne more obstajati vrednost poročen (polje zakonski stan) v povezavi z vrednostjo 10 (polje starost).

SLIKA 8: ODVISNOSTI MED DOSLEDNOSTJO IN DEKLARATIVNIMI OMEJITVAMI TER ŠTEVILOM PONAVLJANJ



Vir: Rudra in Yeo, Key issues in achieving data quality and consistency in data warehousing among large organisations in Australia 1999.

Opravljenih je bilo kar nekaj raziskav med odvisnostjo doslednosti podatkov in ponavljanjem podatkov (redundanca) ter odvisnostjo med doslednostjo in številom deklarativnih omejitev (slika 8) (Rudra in Yeo 1999).

Iz napisanega izhaja, da imata meta podatkovni model ter FIRPM velik vpliv na doslednost podatkov.

Želja po doseganju večje doslednosti podatkov pa ima tudi svoje posledice, ki se kažejo v zmanjšani prilagodljivosti, zato je treba pazljivo premisliti ter poiskati ustrezno ravnotežje med doslednostjo in prilagodljivostjo.

4.3.3 Popolnost

Wang in Strongova (1996) opredeljujeta razsežnost »popolnost« kot podatkovno sposobnost zadovoljevanja potreb za določeno problemske razmere, ki jo Pipino, Lee in Wang (2002) proučujejo s treh različnih vidikov. Po njihovem mnenju je najprej treba zagotoviti popolnost v podatkovni shemi, tj. ali imamo pravilno opredeljene entitete ali imajo entitete opredeljene ustrezne lastnosti. Popolnost se mora zagotoviti tudi za posamezno lastnost z vidika vrednosti, ki jo ta lastnost lahko zavzame, tj. ali imamo prazne vrednosti v stolpcih ali ne. Navedeni vidik popolnosti ustreza stolpčni integriteti v relacijskem podatkovnem modelu in jo ustrezen FIRPM

lahko zagotovi (deklarativne omejitve). Zadnji vidik popolnosti pa je popolnost populacije (populacija omrežnih skupin v Sloveniji šteje 6, in če bi bilo v sistemu/preglednici le 5 elementov, potem populacija ne bi bila popolna). Za vse tri vidike nepopolnosti izračunavamo količnike nepopolnosti.

PREGLEDNICA 8: POPOLNOST Z VIDIKA VREDNOSTI NULL

Vpisna številka študenta	Priimek	Ime	Elektronski naslov	Razlaga vrednosti NULL
12345	Novak	Janez	Janez.Novak@siol.net	
54321	Ahačič	Miha	Null	Ne obstaja
21345	Dolinar	Anže	Null	Obstaja, vendar ga ne poznamo
32145	Perme	Boris	Null	Ne vemo, ali obstaja

V literaturi se pojavljajo še druge opredelitve popolnosti podatkov. Redman (1996) popolnost opredeljuje kot stopnjo pojavnosti podatkovnih vrednosti v zbirkah podatkov. Podobno opredelitev podajajo tudi Jarke, Jeusfeld, Quix in Vassiliadis (1999).

Popolnost nam torej pove, kako nam podatki pomagajo reševati problemsko situacijo. Če bi v združbi radi razposlali obvestilo o novi storitvi obstoječim kupcem, v zbirki podatkov pa nam manjka ogromno naslovov (vrednost lastnosti naslov je NULL), potem so podatki nepopolni.

Batini in Scannapieca (2006) posvečata posebno pozornost popolnosti v relacijskem podatkovnem modelu, kar spet potrjuje primernost teme, ki jo obravnava knjiga. Popolnost v relacijskem podatkovnem modelu proučujeta z vidika prisotnosti/odsotnosti vrednosti NULL ter z vidika domnevanja o popolnosti podatkov.

Vrednost NULL nekega atributa v relacijskem podatkovnemu modelu pomeni neznano vrednost. To ni prazna vrednost, ampak neznana vrednost, kar pomeni, da v realnem svetu ta vrednost najbrž obstaja, vendar je iz različnih razlogov nimamo shranjene v zbirki podatkov. Za opredelitev razsežnosti popolnost kakovosti podatkov je pomembno razumeti, zakaj imamo v zbirki podatkov vrednosti NULL: vrednost v realnem svetu morda ne obstaja ali pa obstaja, vendar nam ni poznana, oziroma obstaja, vendar ne vemo, da obstaja.

Pomemben pa je tudi vidik domnevanja o popolnosti podatkov, pri čemer lahko domnevamo, da imamo v zbirki podatkov vse vrednosti, ki odsevajo stvarni svet (*Closed World Assumption – CWA*), ali pa ne moremo opredeliti, ali so pojavne vrednosti v zbirki podatkov popolne ali nepopolne (*Open World Assumption – OWA*).

Pomembnost obeh vidikov domnevanja poudarjajo Naumann, Freytag in Lester (2004) ter Batini in Scannapieca (2006). Popolnost lahko izrazimo s količnikom kot razmerje med podatki, navedenimi v zbirki podatkov, in številom vseh ustreznih podatkov, ki bi morali biti v zbirki podatkov. Popolnost je spet treba proučiti z vidika nameravane uporabe.

Najlažje je to ponazoriti s primerom. Pogosto se zgodi, da vnašalci niso popolnoma natančni ali vestni pri vnašanju podatkov v informacijski sistem. Recimo, da se v združbi uporablja sistem za spremljanje dela pri projektih, v katerega zaposleni vnašajo čas, porabljen za dejavnosti pri določenem projektu. Seveda zaposleni niso vedno dovolj vestni, da bi vnesli vse podatke. Takšen sistem je dovolj kakovosten z vidika popolnosti za splošen pregled nad porabljenim časom pri projektu, medtem ko je nekakovosten, kadar delo zaračunavamo naročniku.

4.3.4 Zaupanje v podatke

Podatki so lahko v skladu z vsemi razsežnostmi kakovostni, vendar če jim uporabniki ne zaupajo, potem je vse zaman. Nezaupanje je navadno posledica napak v preteklosti, ki so omajale zaupanje v podatke. Po Pradhanovem (2005) mnenju je zaupanju kot razsežnosti kakovosti podatkov namenjeno premalo pozornosti predvsem iz dveh razlogov: če zaupanje obravnavamo kot lastnost uporabnikov, potem je to stvar psihologije, in ne informacijske znanosti, poleg tega pa nekateri avtorji, na primer Naumann in Roth (2004), obravnavajo zaupanje kot podmnožico razsežnosti natančnosti. Wang in Strong (1996) opredeljujeta zaupanje v podatke kot raven sprejemanja, da so podatki resnični, natančni in verodostojni.

Za razumevanje zaupanja je treba proučiti zaupanje z vidika lastnosti končnih uporabnikov ter kot podmnožico natančnosti.

Pradhan (2005) se sprašuje, zakaj pravzaprav potrebujemo razsežnost »zaupanje«. Če ima podatek razsežnost natančnost, potem naj bi se tem podatkom zaupalo, in nasprotno. Podatkom, ki nimajo razsežnosti natančnost, naj se ne bi zaupalo. Vendar se ne ustavi tukaj in postavi še drznejšo trditev. V svojem delu ugotavlja, da je zaupanje osrednja razsežnost kakovosti podatkov in da je natančnost le njena lastnost, ter s tem ovrže trditve, da je zaupanje lastnost natančnosti. Po njegovem mnenju so podatki natančni takrat, ko verjamemo, da so natančni. To dokazuje tudi s primerom zalog v skladišču. Če želimo izvedeti, koliko kosov nekega izdelka imamo v skladišču, bomo opravili poizvedbo v zbirki podatkov. Po izpisu iz informacijskega sistema bi se odpravili v skladišče in prešteli število izdelkov, ki nas zanimajo. Lahko se izkaže, da sta podatka različna, vendar ali to dejansko pomeni, da je vrednost iz sistema napačna. Kaj pa če zaposleni v skladišču hranijo tudi svoje blago, ki ne pripada združbi in ga ne moremo šteti v zalogo. Vidimo torej, da navedeni primer potrjuje, da nam preverjanje blaga v skladišču ne more potrditi natančnosti podatka, torej je od zaupanja odvisno, ali bomo

podatek iz informacijskega sistema šteli kot natančen ali ne. In če pomislimo, tako dejansko delujemo tudi v vsakdanjem življenju. Če v podatek verjamemo, je za nas natančen, če vanj ne verjamemo, podatek za nas ni natančen. Iz navedenega lahko sklepamo, da obstaja razkorak med željo po določenem podatku in možnostjo preverjanja tega podatka, ki izpostavlja zaupanje kot osrednjo razsežnost kakovosti podatkov.

Drugi vidik zaupanja pa je odvisen od uporabnika podatka, torej ko govorimo o zaupanju, vedno mislimo na raven zaupanja določenega uporabnika, pri čemer še vedno ni raziskano, katere lastnosti mora podatek vsebovati, da ga lahko opredelimo kot zaupanja vreden. Pri tem ne mislimo na različne subjektivne lastnosti, ampak na stalne, torej tiste lastnosti, ki so enake za vse uporabnike.

Omenjeni vidik zaupanja, torej zaupanje kot odvisna spremenljivka uporabnika, lahko preverimo tudi z »zdravo« pametjo, kjer premislimo utemeljitve za in utemeljitve proti, ali naj bi določenemu podatku verjeli, pri čemer se osredotočimo na vsebino podatka in proces nastanka podatka. Ko govorimo o procesu nastanka, se neizogibno srečamo s pojmom »*garbage in – garbage out*«. Če je vhod v proces nekakovosten, bo tudi rezultat nekakovosten. Naumann in Roth (2004) sta po analogiji postavila trditev »*quality in – quality out*«, torej če imamo kakovosten vhod v informacijski sistem, bomo dobili tudi kakovosten rezultat. Pradhan (2005) pa dodaja še tretji vidik, in sicer »*quality in – garbage out*«, ko informacijski proces pokvari kakovosten vhod. Vidimo lahko, da proces nastanka podatka, informacije in znanja pomembno vpliva na raven zaupanja kot razsežnost kakovosti podatka.

Kakor koli gledamo na zaupanje kot razsežnost kakovosti podatka, ne moremo izključiti človeške subjektivnosti in procesa sprejemanja odločitev. Uporabnik se mora dejansko odločiti, ali podatkom verjame ali ne, in s tem opredeli tudi natančnost podatkov. V poglavju o odločanju je že bilo omenjeno, da se človek lahko odloča racionalno-analitično, intuitivno in preudarno. Ne glede na način odločanja njegove pretekle izkušnje vplivajo na odločanje, zato tudi pretekle izkušnje s podatki določajo raven zaupanja v sedanje in prihodnje podatke. In ko uporabnik zaupanje izgubi, ga težko dobi nazaj – in v tem primeru so podatki z vidika zaupanja v zbirki podatkov nekakovostni. Ni pomembno, ali na zaupanje gledamo kot na podmnožico natančnosti ali nasprotno, zaupanje je pomembna razsežnost kakovosti podatkov.

4.3.5 Pravočasnost

Pravočasnost spada v časovno razsežnost kakovosti podatkov, kamor nekateri avtorji (Scannapieca, Missier in Batini 2005) prištevajo tudi veljavnost in spremenljivost. Slednjih dveh Wang in Strongova (1996) nista opredelila kot razsežnosti kakovosti podatkov. Ballou (1982,

1985, 1987) pravočasnost opredeljuje kot nezastarelost podatka, čeprav bi slednjo opredelitev hitreje pripisali veljavnosti.

Veljavnost je razsežnost, ki opredeljuje pogostost sprememb določenega podatka, tj. kako pogosto je določen podatek spreminjan. To lahko ponazorimo s primerom podatka o predelavi določenega filma. Film Avtoštopar je nastal leta 1996, v letu 2007 je doživel predelavo. In če v zbirki podatkov nismo posodobili tega podatka, potem je podatek z vidika razsežnosti »veljavnost« z letom 2007 postal nekakovosten. Veljavnost merimo glede na zadnjo spremembo podatka; če je spremenljivost različna (podatek se spreminja hitreje/počasneje v določenem obdobju), navadno merimo povprečno veljavnost.

Spremenljivost nam pove, kako hitro se določen podatek spreminja. Podatek o rojstnem dnevu ima spremenljivost enako 0, saj se ne spreminja, medtem ko ima recimo podatek o zalogi določenega izdelka visoko raven spremenljivosti. Spremenljivost izrazimo s časom, ko ima določen podatek nespremenjeno vrednost.

Pravočasnost pa opredeljuje, ali so veljavni podatki še ustrezni za reševanje določenega problema v določenem trenutku (ali so podatki na voljo, ko jih potrebujemo). Formalnejša opredelitev je nastala v kanadskem statističnem uradu, ki opredeljuje pravočasnost kot čas med zaznano potrebo po podatku in trenutkom, ko podatek postane razpoložljiv (2006). Lahko se zgodi, da imamo v zbirki podatkov povsem veljavne podatke, toda če jih v določenem trenutku iz različnih razlogov nismo uspeli uporabiti, potem so neuporabni. Predpostavimo, da ima združba več enot, ki različno pošiljajo oziroma knjižijo prejete račune in plačila. Nekatere enote zavedejo račune takoj, ko jih prejmejo, nekatere pa čakajo do konca meseca. Če se takšna zbirka podatkov uporabi na primer za pregled odprtih postavk določenega kupca na določen dan, potem je z vidika pravočasnosti podatkovna zbirka nekakovostna. Za pregled trenda naraščanja ali padanja terjatev do kupcev pa je zbirka podatkov kakovostna.

Pravočasnost izračunamo kot $\max(0,1-\frac{veljavnost}{spremenljivost})$ in lahko zavzame vrednosti med 0 in 1, kjer 0 pomeni slabo pravočasnost, 1 pa odlično pravočasnost (Ballou et al. 1998).

Če premislimo o navedenih treh časovnih razsežnostih, lahko opazimo, da imata Wang in Strongova (1996) prav, ko govorita le o eni časovni razsežnosti, tj. pravočasnost, medtem ko je veljavnost razsežnost, ki bi jo lahko opredelili kot lastnost razsežnosti natančnost, razsežnost spremenljivost pa za odločanje v nekem trenutku nima večje vrednosti. Slednji dve trditvi je najlažje obrazložiti s praktičnim primerom. Predpostavimo, da imamo v združbi podatke o cenah izdelkov različnih dobaviteljev, ki se v času spreminjajo. Ko se odločamo za nakup določenega izdelka, nas veljavnost pravzaprav ne zanima, pomembno je samo, ali podatku verjamemo ali

ne, torej ali je podatek natančen. Veljavnost nas pravzaprav le opozori, ali je podatek natančen ali ne, kar opredeljuje kakovost podatkov z vidika razsežnosti natančnost.

Še manjšo vlogo v tem primeru ima spremenljivost, saj je za nas, ko se moramo odločiti v danem trenutku in ne v nekem daljšem časovnem obdobju, pravzaprav popolnoma nepomembno, ali se podatek spreminja hitro ali počasi. Če pa bi odločitev sprejemali v nekem daljšem časovnem obdobju, bi spremenljivost imela večjo težo.

Glede na navedena dejstva lahko rečemo, da edina pomembna časovno pogojena razsežnost kakovosti podatkov je pravočasnost.

4.3.6 Ustreznost podatkov

Ustreznost kakovosti podatkov opredelimo kot zmožnost zadovoljevanja potreb uporabnika podatkov (Brackstone 1999), Wang in Strongova (1996) pa z razsežnostjo ustreznost opisujeta podatke, ki so uporabni in koristni za določeno problemsko situacijo.

Ustreznost lahko preverimo tudi z vidika nameravane uporabe. Predstavljajmo si združbo, ki ima vzpostavljeno zbirko podatkov o materialih, pri čemer ji enak material lahko dostavlja več dobaviteljev. V zbirki podatkov ni zapisa, kateri kos materiala je dostavil določen dobaviteljev. Podatkovna zbirka je z vidika natančnosti in pravočasnosti lahko zelo kakovostna, vendar pa ni kakovostna z vidika ustreznosti. Predpostavljamo namreč, da določen dobavitelj sporoči, da je v dostavljenih materialih nastala serijska napaka, naša združba pa se ne more ustrezno odzvati, saj je po tem, ko je prevzela material, izgubila sled za njegovim dobaviteljem.

4.3.7 Razumljivost podatkov

Čeprav podatki zadostijo vsem predhodnim merilom/razsežnostim, so lahko z vidika razumljivosti popolnoma nekakovostni. V združbi je popolnoma normalno, da se določen račun ali naročilo prekliče zaradi morebitne napake na listini in se izda nova popravljena listina. Z vidika računovodstva je sistem popolnoma kakovosten, kar pa ne velja recimo za analitika, ki proučuje število in trend naročil, saj ga preklicani računi lahko zmedejo (če recimo gleda porabljene število izdanih listin, mu tudi preklicana listina pomeni naročilo).

4.3.8 Dostopnost

Če so podatki dostopni oziroma jih lahko hitro dobimo, potem so po mnenju Wanga in Strongove (1996) podatki kakovostni. Napačno je mišljenje, da z objavo podatkov na svetovnem spletu končamo svoje delo v smislu omogočanja dostopa do podatkov. Razumeti je treba, kaj posameznik potrebuje v navedenem primeru, da lahko dostopa do podatkov (imeti mora računalnik, priključen v svetovni splet, znajti se mora z ustreznim odjemalcem in pravilno

razlagati podatke). Dostopnost torej pomeni oziroma opredeljuje sposobnost končnega uporabnika, da v skladu s svojo kulturo, fizičnim stanjem in tehnološkimi zmožnostmi dostopa do podatkov (Batini in Scannapieca 2006, 34). Predvsem je pomembno, da se dostop do takšnih podatkov zagotovi tudi prizadetim osebam, čemur je veliko pozornosti namenil *World Wide Web Consortium* (2007), ki priporoča, da se mora dostopnost do podatkov prilagoditi naslednjim osebam:

- osebam, ki se ne morejo gibati, videti ali slišati;
- osebam, ki imajo težave z branjem ali razumevanjem besedila;
- osebam, ki niso zmožne uporabljati tipkovnice ali miške;
- osebam, ki imajo manjši zaslon ter počasnejšo povezavo v svetovni splet;
- osebam, ki niso sposobne tekoče govoriti ali razumeti naravnega jezika.

Zaradi navedenih smernic so se razvile številne tehnologije, ki omogočajo predstavitev vsebine na različne načine, na primer sintetizacija govora, Braillova pisava, predstavitev znakovnega besedila z različnimi grafičnimi simboli.
Številne države so sprejele različne ukrepe za zmanjšanje »prepada« med prizadetimi osebami in drugimi ljudmi.

Združbe so uspešne pri zbiranju in hranjenju podatkov, vendar imajo težave pri omogočanju dostopa do shranjenih podatkov (Chung 2002). Navadno so podatki ekskluzivno namenjeni določeni enoti v združbi in druge enote težko ali pa sploh ne morejo dobiti drugih podatkov, razen lastnih, kar je posledica organizacijskih predpisov ali/in zasnove poslovanja. Nedostopnost lahko povzroči resno škodo v združbi. Predstavljajmo si združbo, kjer ima skupina posameznikov podatke o lojalnosti strank, hkrati pa se druga skupina neodvisno loti zbirati takšne podatke, ne vedoč, da ti že obstajajo. Namesto sodelovanja imamo v združbi dve neodvisni skupini, ki bi s sodelovanjem dosegali sinergijske učinke.

Toda želja posamezne združbe po veliki dostopnosti ima tudi negativne posledice. Mullins (2006) opozarja pred velikodušnostjo združb, da omogočijo dostop do podatkov širši množici ljudi. Prav nobenega razloga ni, da bi bili vsi podatki dostopni vsem zaposlenim. Takšno ravnanje združb vzbuja vtis o nezainteresiranosti združb za primerno zaščito svojih podatkov. Vsak pomemben podatek bi moral biti ustrezno varovan in posredovan samo pooblaščenim osebam. Slabo oziroma pomanjkljivo varovanje je tudi najpogostejši vzrok za različne vdore v zbirke podatkov posameznih združb.

Poleg vse naštetih razsežnosti kakovosti podatkov ne smemo zanemariti dostopnosti, čeprav se zdi, da obravnavana razsežnost nima velike vloge pri celotni opredelitvi kakovosti podatkov. Dostopnost izračunamo kot max(0, 1).

4.3.9 Druge razsežnosti

Poleg naštetih razsežnosti obstajajo tudi druge razsežnosti kakovosti podatkov, ki pa jih literatura ne obravnava tako pogosto kot že navedene razsežnosti kakovosti podatkov. Omenil sem že, da ne obstaja enotno mnenje o razvrščanju razsežnosti po pomembnosti, vendar je obstoječa razdelitev dokaj smiselna. Če izhajamo iz teorije odločanja, lahko hitro in brez slabe vesti opredelimo najpomembnejše razsežnosti. V tem primeru je pomembno, da imamo pravočasno pri roki natančne podatke, ki jim zaupamo. Torej so natančnost, pravočasnost, dostopnost in zaupanje bistvene razsežnosti.

PREGLEDNICA 9: DRUGE RAZSEŽNOSTI, KOT STA JIH OPREDELILA WANG IN STRONGOVA

Razsežnost	Kaj opredeljuje?
Objektivnost	Nepristranskost podatkov
Ugled	Obsežnost spoštovanja oz. ugleda podatkov z vidika vsebine in vira
Razlagalnost	Obseg podatkov v ustreznem jeziku in merskih enotah
Razumljivost	Obseg nedvoumnih in enostavno razumljivih podatkov
Jedrnatost	Obseg jedrnatosti predstavitve podatkov (podatki ne smejo biti preobsežni)
Količina	Mero za pravo količino podatkov v smislu ustreznosti

Vir: Wang in Strong, Beyond accuracy: what data quality means to data consumers, 1996.

Ko pri obravnavanju razsežnosti kakovosti podatkov upoštevamo namen uporabe podatkov, namen uporabe lahko ocenimo kot pomembno merilo za ocenjevanje kakovosti podatkov, zato bi bilo idealno popisati najprej vse zahteve uporabnikov in slednjim prilagoditi informacijski sistem ter zbirko podatkov. Vendar je resnica daleč od ideala, saj se pogosto zgodi, da se informacijski sistem uporablja za namene, za katere ni bil načrtovan, in sicer zaradi:

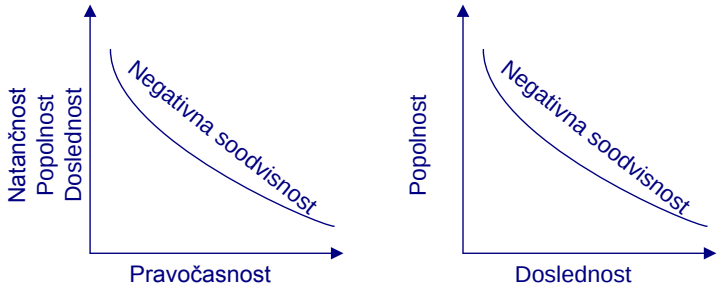
- nove zakonodaje,
- nastopa na novem tržnem segmentu,
- prevzema združb,
- rasti združbe in novih zahtev.

4.3.10 Soodvisnost razsežnosti

Razsežnosti kakovosti podatkov pa med seboj niso neodvisne, ampak so soodvisne. Poudarjanje ali iskanje določene razsežnosti lahko negativno vpliva na soodvisne razsežnosti. Očitna soodvisnost obstaja med natančnostjo, popolnostjo, doslednostjo in pravočasnostjo. Če hočemo

doseči večjo natančnost, popolnost in doslednost, potrebujemo več »varnostnih mehanizmov« ter preverjanj. Postopki preverjanja potrebujejo čas in negativno vplivajo na pravočasnost. Obstajajo področja, kjer je pravočasnost pomembnejša, in področja, kjer je pravočasnost manj pomembna. Predstavljajmo si spletno stran s ponudbami »*last minute*«. Za lastnike teh spletnih strani je pravočasnost izjemno pomembna, saj so potovanja »*last minute*« na voljo le nekaj dni ali ur pred začetkom potovanja (mislimo na resnične ponudbe »last minute«, in ne na ponudbe, kjer potovanja za mesec september lahko rezervirate že meseca junija, čeprav ima oznako »last minute«). V naglici se lahko pojavijo in se pojavljajo napake, na primer napačne cene potovanja, napačni hotel ipd., vendar pravočasnost odtehta nenatančnost, saj bi bil podatek popolnoma nekoristen, če bi bil na voljo šele po začetku potovanja. Ko govorimo o preverjanju in pošiljanju vavčerjev, pa natančnost prevlada nad pravočasnostjo (rahle zamude) in velikokrat se zgodi, da vavčer dobimo šele teden dni pred odhodom, in ne nekaj tednov prej, kot nam je obljubljeno.

SLIKA 9: SOODVISNOST MED RAZLIČNIMI RAZSEŽNOSTMI



Zelo očitna je tudi odvisnost med doslednostjo in popolnostjo. V tem primeru izhajamo iz vprašanja, ali je boljše, da imamo manj podatkov, ki so bolj dosledni, ali da imamo več podatkov, vendar je doslednost manjša (Ballou in Pazer 2003 ; Sadhegi in Clayton 2002). Odgovor je spet odvisen od narave problema oziroma nameravane uporabe, o kateri podrobno govori Olson (2003). Pri izvajanju analiz z različnimi statističnimi tehnikami je za nas popolnost pomembnejša kot doslednost, ki jo v tem primeru lahko zagotovimo z različnimi postopki.

Ko pa razmišljamo o javni objavi ocen študentov pri nekem izpitu, je pomembnejša doslednost, saj v nasprotnem primeru lahko povzročimo veliko škode oziroma čustvenih pretresov. V tem primeru lahko izpustitev nekega študenta s seznama (četudi je to nedopustno) hitro popravimo, medtem ko napačne objave ocen ni tako lahko popraviti, ker so po navadi vpletena čustva.

4.4 Nadgrajena matrika kakovosti podatkov

Opisana opredelitev kakovosti podatkov ima številne kritike, ki opozarjajo na problem statičnosti in premajhne prilagodljivosti, ki izhaja iz navedenih opredelitev. Avtorja Dasu in Johnson (2003) opozarjata na potrebo po novih opredelitvah, ki bi odsevale značilnosti

današnjega časa. Podatki so doživeli velike vsebinske in količinske spremembe, ki so spremenili tudi način ravnanja s podatki. Če smo včasih ravnali predvsem s podatki iz znanih izvorov oziroma virov podatkov in poznane vsebine, so združbe v sedanjosti mnogokrat soočene z neznanimi podatki, ki so proizvodni učinek avtomatskega procesa, oblika (format) je pogosto neznana, podatke črpamo iz različnih virov v združbi in tudi iz okolja, pri združevanju podatkov iz različnih virov pogosto manjkajo ključi, ki so nujni za uspešno združitev. Pri podatkih pa se je spremenil tudi vidik končne uporabe. Zaradi omejene zmogljivosti računalniške opreme v preteklosti so združbe podatke zbirale s točno določenim namenom (niso jih zbirale »na zalogo«), temu primerna je bila tudi njihova obdelava pred vpisom v skladišče podatkov. Cene strojne računalniške opreme so v zadnjih letih močno padle, zmogljivost pa je narasla, zato se podatki pogosto zbirajo in shranjujejo z odsotnostjo vizije njihove uporabe, tj. »na zalogo«. Navedene značilnosti podatkov so po mnenju Dasuja in Johnsona (2003) zadosten in potreben razlog, da začnemo razmišljati o novi opredelitvi izraza »kakovost podatkov«, ki ga obravnavata z vidika celotnega procesa zagotavljanja kakovosti podatkov (angl. *data quality continuum*). Proces zagotavljanja kakovosti podatkov razdelita na zbiranje in posredovanje podatkov, povezovanje in shranjevanje, vnovično pridobivanje ter analizo podatkov.

4.4.1 Zbiranje in posredovanje podatkov

Zbiranje podatkov je začetek procesa ravnanja s kakovostjo podatkov, čeprav ga včasih »prehiti« proces načrtovanja zbiranja podatkov, v katerem se odločamo o količini in vrsti podatkov, ki jih bomo zbirali.

V procesu zbiranja največkrat naletimo na probleme napačnih (nenatančnih, nedoslednih in nepopolnih) vnosov v zbirko podatkov (na primer namesto 17 vnesemo 71; kraj izdaje računa ima vrednost NULL, Ljubljana, Lj. …), probleme podvojenih oziroma ponavljajočih se vrednosti podatkov, lahko pa naletimo na povsem vsebinske težave. Slednje so največkrat posledica pomanjkljivega načrtovanja, kjer določimo količino in vrsto podatkov, ki jih bomo zbirali. Če zbiramo podatke po tednih, in ne dnevih, analitična služba pa želi spremljati dogodke po dnevih, potem obstaja razkorak med željo in zmožnostmi, ki ga ni preprosto odpraviti, saj je treba spremeniti postopke za zajemanje podatkov.

V procesu zbiranja podatkov obstajajo številne možnosti za zmanjšanje napak:

- načrtovanje procesa zbiranja podatkov v smeri avtomatizacije, ki zmanjšuje potrebno število ročnih posegov v podatke;
- primerna zasnova uporabniških vmesnikov (vnosne maske, vrste vnosnih polj);
- zasnova ustreznega meta podatkovnega modela;
- razdelitev odgovornosti za kakovost podatkov posameznikom;

- pogoste revizije procesa ravnanja s kakovostjo podatkov.

Napake pa se lahko pojavljajo tudi v procesu posredovanja podatkov, tj. prenos podatkov iz vira podatkov v trajnejšo obliko, in sicer zbirko podatkov. Pri prenosu lahko podatki izgubijo del svoje vsebine (okrnitev) ali pa se podatki celo izgubijo. Zakaj se pojavljajo takšni problemi? Podatke zaradi različnih vzrokov pred shranjevanjem v zbirko podatkov obdelamo (združevanje/agregiranje, ureditev privzetih vrednosti), s čimer okrnemo podatke z vidika podrobnosti, prav tako pa je pri takšnem ravnanju s podatki mogoča količinska izguba podatkov. Da bi preprečili okrnitev in izgubo podatkov, moramo uporabljati transakcije, ki preprečujejo, da bi se del podatkov prenesel, drugi del pa ne, in spremljati povratno informacijo o izvršitvi/neizvršitvi transakcije. Dasu in Johnson (2003) naštevata še druge možnosti, ki pa kažejo na njuno nepoznavanje sodobnih SUBP-ov (na primer *DTS, Integration Services*), ki avtomatsko omogočajo preverjanje različnih odvisnosti podatkov ter izpeljavo prenosa znotraj transakcije. V opravičilo jima lahko štejemo, da je bilo njuno delo izdano leta 2003, ko nekatera orodja (na primer MS SQL 2005) še niso obstajala na trgu.

4.4.2 Nadzorovanje podatkov

Kljub moderni in zanesljivi tehnologiji, ki je prenos in ravnanje s podatki spremenila v obvladljiv proces, se pri ravnanju s podatki še zmeraj pojavljajo napake, zato je nadzor oziroma preverjanje podatkov nujno potrebno opravilo, če hočemo slediti cilju boljše kakovosti podatkov. Moderna orodja nam omogočajo učinkovito sledenje podatkom (omogočajo vodoravni in navpični prerez podatkov, na primer *Data Views v Integrations Services*), hkrati pa učinkovito in pregledno določajo pravice dostopa do kontrolnih podatkov. Omogočiti dostop do kontrolnih podatkov pooblaščenim osebam je pomembna dejavnost, saj so pooblaščene osebe po navadi vsebinski poznavalci podatkov in hitro odkrijejo napake. Poleg ročnih posegov moderna orodja poznajo posebne dogodke, alarme (angl. *alerts*), ki lahko delujejo kot sprožilno dejanje za določeno akcijo popravljanja podatkov.

4.4.3 Shranjevanje podatkov

Podatke je treba shraniti na primerno mesto v primerno strukturo in zagotoviti ustrezno varnost. Bolj priljubljen sistem za shranjevanje podatkov je v sodobnem času relacijski SUBP, ki zagotavlja visoko raven varnosti s stališča transakcij, varnostnih izvodov (v nadaljevanju kopije) ter dostopa pooblaščenih oseb.

Transakcije so osnovni dogodki, ki poskrbijo, da se izvedejo vsi ukazi (na primer prenesejo in shranijo se vsi podatki) v posamezni transakciji ali pa nobeden, hkrati pa sporočijo tudi status o

svoji uspešnosti oziroma neuspešnosti. Uporaba transakcij zahteva določeno znanje, ki pa ga po mojem mnenju skrbniki zbirk podatkov nimajo dovolj oziroma je pomanjkljivo.

Varnostne kopije so izredno pomembne, saj zagotavljajo nemoteno delo v primeru različnih okvar delovnih zbirk podatkov. Vendar tudi načrtovanje varnostnih kopij ni tako preprosto dejanje, kot se morda zdi na prvi pogled. Odločiti se moramo za vrsto varnostne kopije (na primer v MS SQL imamo možnost *full backup*, *diferential backup* ali *transactional backup*) ter določiti časovni razmik med posameznimi varnostnimi kopijami; krajši časovni razmik je boljši z vidika varnosti, a slabši z vidika učinkovitosti. Varnostno kopijo celotne zbirke podatkov po navadi delamo čez noč, v dnevnem času pa se moramo zadovoljiti z varnostnimi kopijami sprememb med posameznimi varnostnimi kopijami. Problem nastane, ko je podatkovna zbirka prevelika, da bi uspeli narediti celotno varnostno kopijo. Takrat moramo uporabiti druge strategije varnostnih kopij (glej npr. Grassi 2007 ; Monahan 2006 ; Rusell 2006).

Ko govorimo o varnosti, je treba upoštevati več vidikov. Vsako varovanje se najprej začne s fizičnem varovanjem, torej s preprečitvijo dostopa nepooblaščenim osebam do zbirke podatkov. Varovanje se nadaljuje z zaščito strežnika ter zbirke podatkov. Ne smemo pozabiti, da je shranjevanje varnostnih kopij na istem strežniku oziroma celo na istem fizičnem mestu (v isti zgradbi) neprimerno, še posebno če gre za zelo občutljive podatke in ob upoštevanju najbolj črnega scenarija, na primer naravnih nesreč (glej npr. Litwin 2004 ; Schweitzer 2004).

4.4.4 Povezovanje podatkov

Globalizacija in vedno večja konkurenca močno vplivata na način poslovanja združb. Združbe iz različnih razlogov odpirajo poslovne enote na razpršenih krajih, ki imajo po navadi tudi svoje zbirke podatkov, ali pa delujejo prek oddaljene povezave oziroma uporabljajo obe možnosti (privzeto delajo prek oddaljene povezave, ob izpadu povezave začnejo uporabljati lokalno zbirko podatkov, ki se po vnovični vzpostavitvi povezave uskladi z osrednjo zbirko podatkov). Včasih pa je delo na območni (v nadaljevanju lokalni) zbirki podatkov in poznejše usklajevanje (na primer replikacije) edini mogoči način dela, saj nekateri informacijski sistemi »ne zmorejo« veliko istočasnih dostopov in jih je treba razbiti na manjše enote.

Kadar koli imamo več virov podatkov, je ne glede na način dela nujno povezovati podatke zaradi potrebe po splošnem pregledu managementa nad poslovanjem združbe in različnih analiz. Pri povezovanju podatkov lahko naletimo na težave, še posebno če enote združbe niso delovale usklajeno pri osnovnih nastavitvah PIS. Kako združiti podatke, če nimamo skupnega imenovalca oziroma ključa, po katerem bi podatke združevali? Pomislimo na združbo, katere enote imajo popolno prostost pri vnosu šifrantov kupcev, dobaviteljev ter poslovnih učinkov in ima isti kupec, ista stranka ter isti poslovni učinek drugačen glavni ključ v vsaki poslovalnici. V tem

primeru je treba podatke »prevesti« na skupni imenovalec s pomočjo lastnosti/polja, ki jih enolično določa, torej polja, ki je kandidat za glavni ključ (kupci – davčna številka, dobavitelji – matična številka, poslovni učinek – klasifikator). Če takšne lastnosti v zbirki podatkov nimamo, nas čaka naporno delo pregledovanja podatkov in vnos »umetnega« podatka. V znani slovenski združbi, ki ima 5 poslovnih enot, smo pri kupcih uvedli novo polje, ki smo ga poimenovali »super šifra stranke« in je imelo edinstveno vrednost znotraj celotne skupine združb. 5 oseb je tri tedne pregledovalo podatke in jim poskušalo na podlagi naziva ter drugih lastnosti pripisati vrednost, ki je omogočala povezovanje podatkov znotraj skupine združb ter izdelavo skladišča podatkov, nadgrajenega s tehnologijo OLAP.

Pri povezovanju pa naletimo tudi na drugačne težave. Velikokrat se zgodi, da ima posamezna poslovna enota združbe drugačno strukturo podatkov kot ostale enote. V tem primeru je povezovanje bolj zapleteno, saj zahteva poznavanje različnih struktur podatkov. Obstoječo težavo se da omiliti s kakovostnim meta podatkovnim modelom, ki omogoča hitro povezovanje istopomenski polj, ter s pretvorbo lastnosti na skupni imenovalec. Sodobnejši načini povezovanja podatkov navadno potekajo s pomočjo XML, ki poleg vrednosti vsebuje tudi meta podatke, kar močno olajša delo, a hkrati negativno vpliva na učinkovitost procesa povezovanja z vidika hitrosti.

Težave pri povezovanju podatkov se največkrat pojavijo, ko združba prehaja z informacijskega sistema enega proizvajalca na informacijski sistem drugega proizvajalca. V tem procesu prvi ni zainteresiran za sodelovanje in pomoč pri povezovanju, zato ima »novi« ponudnik navadno zahtevno delo. Navedene vzroke Dasu in Johnson (2003) poimenujeta kot politične in sociološke vzroke. Neusklajeni podatkovni modeli so po navadi tudi vzrok za težave pri replikacijah podatkov, čeprav so sodobni SUBP že močno poenostavili ravnanje z replikacijami.

Pri povezovanju podatkov je pomembno njegovo načrtovanje, ki nam prihrani marsikatero težavo. Načrtovanje ustrezne strojne računalniške in programske opreme je zelo pomembno, zato da se ne odločimo za premalo zmogljive računalniške sestavine, kajti poznejši nakup in zamenjava z drugo, zmogljivejšo opremo, je navadno cenovno manj ugodna.

Poleg načrtovanja pa lahko število težav zmanjšamo tudi s pregledovanjem podatkov ter spoznavanjem njihove vsebine, pri čemer lahko ugotovimo tudi druge lastnosti/polja podatkov, po katerih lahko povezujemo podatke med seboj.

Za podrobnejši opis metod za povezovanje podatkov je na voljo veliko različnih člankov (glej npr. Rittman 2007 ; Torr 2007).

4.4.5 Analize podatkov

Analize so zadnji korak v procesu zagotavljanja kakovosti podatkov. Tradicionalno analize niso vzrok kakovosti podatkov, vendar z vidika širše obravnave (podatki + analize = rezultati) postanejo analize pomemben dejavnik vpliva na kakovost podatkov.

V okviru analiz na kakovost podatkov vpliva predvsem količina podatkov, ki pogosto povzroča potrebo po zmanjšanju količine podatkov, kar dosežemo z izbiro vzorca podatkov iz celotnega nabora podatkov. Takšna metoda je ustrezna, ko nas ne zanimajo podrobnosti ali izjemne vrednosti, temveč nam za rešitev težav zadoščajo agregirani oziroma zgoščeni podatki.

Z razvojem različnih orodij za poslovno obveščanje, kamor prištevamo tudi SOZP, se je pojavilo prepričanje o vsemogočnosti modernih orodij, ki naj bi po mnenju nekaterih popolnoma nadomestile analitike v združbah. Včasih pa je potrebno posebno znanje, ki ga je nemogoče opisati z različnimi modeli in kjer ima analitik še vedno osrednjo vlogo. V želji odkriti zakonitosti »za vsako ceno« uporabniki uporabljajo številne modele v upanju, da bo eden od njih že ustrezal. Takšno ravnanje, v odsotnosti zdravega razuma, lahko ima velike negativne posledice, ki zmanjšujejo zaupanje tudi v druge pravilne in koristne analize.

V vsakem koraku procesa zagotavljanja kakovosti podatkov je treba uporabljati znanje z ustreznega področja. Pomembno je, da ima uporabnik oziroma lastnik podatkov/vira podatkov dovolj ustreznega znanja za izbor primernega modela za analize in da zna pravilno obrazložiti rezultate analiz. S takšnim ravnanjem ima združba veliko možnosti, da si zagotovi kakovostne podatke, ki jih ob ustrezni uporabi uspešno spremeni v koristno znanje.

4.5 Vzroki nekakovostnih podatkov

Raziskava TDWI (podrobneje predstavljena na str. 4-5) iz leta 2005 (ter predhodno iz leta 2001) je pokazala, da težave s kakovostjo podatkov izhajajo iz informatike (IT) in poslovanja združbe (Russom 2006). Težave, ki izhajajo iz informatike, so predvsem težave zaradi posodabljanja rešitev[12] (46 %) ter sistemskih napak (25 %). Po drugi strani pa iz poslovanja izhajajoče težave lahko razdelimo na vnos podatkov (75 %), pričakovanja uporabnikov (40 %) in združek obeh dejavnikov (75 %). Veliko težav nastane tudi zaradi zunanjih dejavnikov, in sicer predvsem zaradi napačnega vnosa podatkov pri strankah, na primer prek svetovnega spleta (26 %). Omenjena raziskava je izpostavila tudi podatke, ki so najbolj občutljivi glede kakovosti. Pri občutljivosti prednjačijo podatki o strankah (74 %), poslovnih učinkih (43 %), finančni podatki (36 %), podatki o prodaji (27 %), podatki iz sistemov ERP (25 %), podatki o zaposlenih (16 %) ter podatki iz

[12] V oklepajih so navedeni deleži anketiranih združb v ZDA, ki so pritrdile posamezni težavi in vzroku.

podružnic – če ima združba več poslovnih enot (12 %). Največja dodana vrednost raziskave pa je v opredelitvi vzrokov za slabo kakovost (padajoča razvrstitev):

- neusklajene opredelitve entitet (75 %),
- napačen vnos podatkov zaposlenih (75 %),
- različni prehodi in obdelave podatkov (46 %),
- različna pričakovanja uporabnikov (40 %),
- zunanji podatki (38 %),
- napačen vnos podatkov pri strankah (26 %),
- sistemske napake (20 %),
- spremembe vhodnih postopkov (20 %),
- drugo (7 %).

Neusklajene opredelitve entitet so skupaj z napačnim vnosom podatkov zaposlenih najpogostejši vzrok za nekakovostne podatke. Ko govorimo o neusklajeni opredelitvi entitet, moramo poudariti dejstvo, da so podatki lahko popolnoma kakovostni (natančni, dosledni), vendar jih uporabniki ne razumejo, saj imajo svoj pogled na entiteto. Težave so še posebno izrazite v združbah, ki imajo različne informacijske sisteme. V oddelku trženja imajo sistem, ki na primer stranko opredeljuje/predstavlja drugače kot informacijski sistem v prodaji. Iz takšnih in podobnih razlogov imajo uporabniki težave pri ugotavljanju, kaj je pravilna opredelitev stranke.
Napačen oziroma nepravilen vnos podatkov zaposlenih je poleg neusklajene opredelitve entitet najpogostejši vzrok za nekakovostne podatke, ki se je pojavil z računalniško podprtimi informacijskimi sistemi in ki nam ga najbrž nikoli ne bo uspelo povsem odpraviti, saj je v središču omenjenega problema človek, ki je zmotljiv. Težave z vnosom se navadno z ustreznim načrtovanjem uporabniškega vmesnika, preverjanjem in čiščenjem podatkov pred vpisom v zbirko podatkov, usposabljanjem in spodbujanjem zaposlenih ter rednimi revizijami uspe omiliti.

Iz navedenega je razvidno, da so podatki o strankah najbolj občutljivi za napake. Ti podatki se tudi najhitreje spreminjajo, na primer naslov stranke, osebe za stike, in ob vsaki spremembi obstaja možnost za napake. Friedman (2006) navaja še druge vzroke, zakaj so podatki o strankah najbolj občutljivi za napake. Podatke o strankah v posamezni združbi vnaša množica različnih ljudi (trženjski oddelek, prodaja, oddelek reklamacij …) in po navadi vsak od njih vnese svojo različico »resnice« (na primer vnos naziva stranke: SAOP, SAOP Računalništvo, SAOP d.o.o …). Friedman (2006) izpostavlja tudi problem različnih informacijskih sistemov v združbi, ki zahtevajo različne načine vnosa podatkov (na primer vrstic za naslov je lahko 1, 2 ali več). Na tem mestu je treba poudariti, da se podatki o strankah po navadi izvažajo in uvažajo tudi v druge informacijske sisteme (na primer replikacije med razpršenimi enotami), kar je idealno, da se nekakovostni podatki hitro razširijo po celotni združbi. Po drugi strani pa takšen način

poslovanja omogoča, da z zagotovitvijo kakovostnih podatkov že pri izvoru uspemo zagotoviti kakovost podatkov v celotni združbi. Friedman (2006) ugotavlja, da večje stranke navadno poslujejo pod različnimi imeni, kar vnaša dodatno zmedo v razumevanje »resnice« in poslabšuje kakovost podatkov (Steklarna Hrastnik, na primer, je združba, ki pod svojim imenom združuje še združbe Opal, Vitrum in Stedek. Pri analizi poslovanja s steklarno Hrastnik je torej treba upoštevati tudi druge identitete, ki jih ta združba ima).

Popolne kakovosti združba ne more doseči, to je dejstvo. Treba se je truditi za kakovostne podatke, ki zadostujejo večini uporabnikov. Olson (2003, 34) ugotavlja, da so zbirke podatkov, ki vsebujejo 0,5 % nenatančnih podatkov, opredeljene kot zelo kakovostne zbirke podatkov. Olson (2003) v svojem delu povzema najpogostejše vzroke za nekakovostne podatke, hkrati pa izpostavi nekakovost podatkovnega modela kot vzrok nekakovostnih podatkov.

V nadaljevanju si bomo podrobneje pogledali štiri vzroke za nastanek nekakovostnih podatkov, in sicer vnos podatkov, staranje, ETL ter uporabo podatkov.

4.5.1 Vnos podatkov

Vnos podatkov je najpogostejši vzrok na nastanek netočnih podatkov, tj. nekakovostnih, podatkov. Osrednjo vlogo v procesu vnašanja podatkov igra človek, ki je zmotljivo bitje. Morda ima slab dan in v svoji površnosti vnese naziv stranke SOAP namesto SAOP, s padajočega seznama barv izbere modro namesto rdeče ali pa vnese pravilen podatek v napačno polje (na primer zamenja polji ime in priimek).

Večina vnosov podatkov se navadno začne z nekim obrazcem, na primer izpolnjen potni nalog na papirju, poročilo o opravljenih urah na terenu, popis osnovnih sredstev na papirnatem obrazcu. Izpolnjevanje obrazcev je pogosto zadnje dejanje nekega opravila, in če je človek pod pritiskom ali preprosto len oziroma ga »administracija ubija«, bo dejavnost izpolnjevanja opravil površno ali pa je sproti sploh ne bo opravil, temveč bo obrazce izpolnjeval kampanjsko, ko se bo že mudilo. Posledica tega je veliko število napak v podatkih. Primer takšnega načina dela sem doživel vsaj na dveh področjih, na področju popisovanja osnovnih sredstev ter pospeševanja prodaje. Predvsem slednje je bilo trn v peti pospeševalcem prodaje, ki so po pregledu polic in razgovoru s poslovodjem navadno kar v avtu ali pa popoldne izpolnjevali obrazce za poročila. Dejavnik, da so poročila pisali nekaj ur po opravljenem delu, se je združil z dejavnikom slabe volje zaradi dodatnega dela v prostem času, kar je pomenilo idealno izhodišče za ogromno napak. Na srečo je z razvojem računalniške in komunikacijske tehnologije takšno delo postalo preteklost, na primer uvedba dlančnikov in sprotno pisanje poročila (še boljše je samodejno zajemanje podatkov, na primer s črtno kodo), in količina nekakovostnih podatkov se je močno zmanjšala (Olson 2003).

Navedene napake pa moramo ločiti od namernih napak pri vnosu podatkov. Namerne napake pri vnosu podatkov po navadi nastanejo zaradi nepoznavanja pravilne vrednosti, želje po anonimnosti oziroma neizdajanju podatkov in zaradi osebnih koristi. Ko govorimo o nepoznavanju vrednosti, navadno mislimo na izpolnjevanje raznih obrazcev, ki niso dovolj jasni in uporabnika zmedejo. Iz sestave obrazca tudi ne izhaja, ali je določen podatek pomemben ali ne, zato uporabnik vnese napačno vrednost ali vrednosti sploh ne vnese. Takšni podatki po navadi niso pomembni za izvedbo določenega opravila, lahko pa so zelo koristni. Podatek o registrski številki vašega osebnega avtomobila pri prijavi v hotelsko sobo se zdi popolnoma nepomemben, toda če vas hoče hotelsko osebje obvestiti o stvareh v zvezi z avtomobilom (na primer prižgani avtomobilski žarometi), je nujen.

V današnjem času svetovnega spleta se je pojavilo ogromno število uporabniških spletnih rešitev oziroma spletnih strani, ki od nas zahtevajo ime, priimek in elektronski naslov. Zaradi večje zavesti o pojavu neželene pošte ter želje po anonimnosti velikokrat vnesemo neresnične podatke, ki navadno ne vplivajo na izvršitev opravila. Takšni podatki postanejo popolnoma neuporabni za analize, zato morajo združbe izbrati pravilno strategijo pri zbiranju takšnih podatkov, uporabnikom morajo v zameno ponuditi neko nagrado. Združba Amazon pri predlaganju nakupov uporablja SOZP in uporabnike ozavešča, da več in boljše podatke bodo vnesli, zanesljivejši bo rezultat Amazonove analize. Veliko napačnih podatkov dobijo tudi združbe, ki v zameno za naše osebne podatke dovolijo uporabo preizkusne različice njihove uporabniške rešitve. Uporabniki vnesejo napačne podatke, saj na spletu poiščejo »*crac*k« in preizkusno različico programa spremenijo v standardno različico ter združbo prikrajšajo za prihodek.

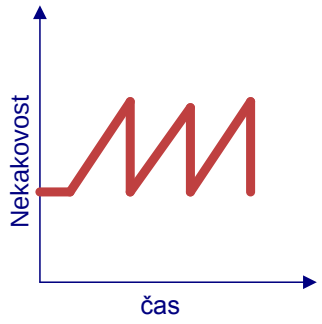
Tretji vzrok namernih vnosov napačnih podatkov je pridobivanje osebne koristi (Olson 2003). Vnašanje poročil o delu na terenu, ki ga zaračunamo strankam, je izkušnja, ki mnogo združb zavede, da dodajo kakšno dodatno uro na poročilo in to dodatno uro stranki zaračunajo. Veliko skušnjav imajo tudi avtomobilski servisi, ki proizvajalcu prikažejo zamenjavo avtomobilskih delov ali avtomobilska popravila kot opravila, ki spadajo v jamstvo, hkrati pa lastniku avtomobila svoje storitve tudi zaračunajo. Plačilo zaposlenim glede na količino vnesenih podatkov, pri čemer kakovosti podatkov ne preverjamo, je tudi velika priložnost za veliko število nekakovostnih podatkov.

4.5.2 Staranje podatkov

Vse se stara, tudi podatki v zbirkah podatkov. Ko podatke vnesemo v zbirko podatkov, so ti po navadi natančni, a natančnost se sčasoma slabša. Pri tem je treba poudariti, da se ne vsi podatki starajo, na primer datum rojstva, EMŠO, spet drugi podatki se močno starajo, na primer stanje zalog blaga. Podatki, ki opredeljujejo entitete oziroma objekte, se ne starajo. Entiteta »študent«

je na primer opredeljena z vpisno številko študenta, imenom in priimkom – ti podatki po navadi ostanejo nespremenjeni, medtem ko nekateri podatki, ki pomenijo lastnosti entitet oziroma objektov, lahko postanejo precej zastareli. Če nadaljujemo s prejšnjim primerom študenta, bi bili takšni lastnosti starost in naslov.

SLIKA 10: NEKAKOVOST V ODVISNOSTI OD ČASA PRI STARAJOČIH SE PODATKIH



Vir: Olson, Data quality the accuracy dimension, 2003, str. 51.

Še posebno problematične so uporabniške rešitve za področje kadrovanja. To je spet povezano z naravo človeka, ki nima občutka dolžnosti za sporočanje spremenjenih podatkov v kadrovsko službo. Obstajajo priporočila, ki se ukvarjajo s tovrstno problematiko. Vsak zaposleni naj bi takoj po vnosu pregledal svoje podatke in potrdil njihovo natančnost. Enak proces naj bi se ponavljal vsako leto.

Značilnost starajočih se podatkov vidimo na sliki 10, ki prikazuje odvisnost med nekakovostjo (razsežnostjo natančnost) in časom.

Združbe na žalost zanemarjajo problem starajočih se podatkov, vendar ne zaradi ekonomskega izračuna (stroški ukvarjanja s tovrstnimi podatki presegako koristi), saj zaposleni o tem preprosto ne razmišljajo. Olson (2003) ugotavlja, da so stroški ravnanja s temi podatki po navadi nizki. Poudarja vlogo meta podatkovnega modela, ki mora nakazovati morebitne težave zaradi staranja podatkov in uporabnike predhodno ustrezno opozoriti.

4.5.3 ETL podatkov

Ko govorimo o ETL (*Extract-Transforming-Loading*), po navadi mislimo na proces izdelave skladišča podatkov. Pri izdelavi skladišča podatkov lahko potegnemo vzporednice s proizvajalnim procesom. Eno od končnih opravil pri proizvajalnem procesu je pregled kakovosti blaga, preden ga odpremimo za kupce. Blago, ki ne ustreza kakovosti, vrnejo v proizvajalni proces z namenom odpraviti napake. Če napak ne zmoremo odpraviti, se blago zavrže. Omenjeni proces kontrole blaga zagotavlja, da za kupce odpremimo kakovostnejše blago, hkrati

pa proces kontrole povišuje stroške proizvodnje. Vendar obstaja drugačna možnost, s katero kakovost »vgradimo« v blago, s čimer se zmanjša potreba po velikem nadzoru kakovosti blaga (Deming 1986). Popolnoma enako je s kakovostjo podatkov v procesu izdelave skladišča podatkov. Silvers (2006) navaja, da je kakovost podatkov v procesu ETL mogoče doseči s kakovostnimi metodami ETL, hkrati pa opozarja na mogoče vidike nekakovosti podatkov v procesu ustvarjanja skladišča podatkov. Ti vidiki so:

- podatkov v skladišču podatkov je premalo ali preveč (manjkajo vrstice ali pa jih je preveč);
- izguba podrobnosti;
- kršenje referenčne integritete;
- zmanjšanje popolnosti kot razsežnost kakovosti podatkov.

Večino naštetih problemov je mogoče z ustrezno metodo rešiti, na primer uporaba stavkov SQL, tujih ključev, ustreznega relacijskega in meta podatkovnega modela, razdelitev odgovornosti za kakovost podatkov oziroma določitev skrbnikov virov podatkov.

Načrtovanje ustreznih postopkov prenosa podatkov iz različnih virov v skladišče podatkov ni enostavna naloga, čeprav so omenjeno nalogo moderna orodja zelo olajšala. Če primerjamo med seboj dva izdelka, in sicer MS SQL 2000 DTS in MS SQL 2005 IS, nam kmalu postane jasno, kako pomemben postaja vidik kakovosti za poslovanje združb. V MS SQL 2005 so že vgrajene metode za iskanje in čiščenje nekakovostnih podatkov, ki morda niso tako zmogljive kot druge metode, vendar MS SQL 2005 zagotavlja celo paleto tesno med seboj povezanih orodij za ravnanje s podatki, kar je za posameznega razvijalca izredno pomembno.

Deming (1986) navaja tudi pomembnost neprestanega izboljševanja produktivnosti in kakovosti, ki posledično znižuje stroške. O podobnosti med Demingovo trditvijo in postopki ETL govori Silvers (2006), ki poudarja vlogo izboljšav obstoječih postopkov ravnanja s podatki, da bi preprečili morebitne napake oziroma nekakovostne podatke. V tem primeru govorimo o preprečevanju namesto o »zdravljenju«.

Veliko težav bi bilo prihranjenih, če bi bili viri podatkov urejeni, skrb za urejanje pa prepuščena ustreznim zaposlenim, vendar Silvers (2006) opozarja na zmotno mišljenje, da bi težave s kakovostjo podatkov popolnoma odpravili. Jasno pove, da se pri podatkih »čudeži ne dogajajo« in da bodo težave nastopile, vprašanje je le, kako velike bodo. Poslovanje združb je dinamično, spreminjajo se vhodne maske, dobavitelji in njihovi podatki prihajajo ter odhajajo. Pri načrtovanju postopkov ETL moramo upoštevati vse navedene težave, čeprav v trenutku nastajanja in preizkušanja postopkov ETL težave še ne obstajajo. Silvers (2006) poudarja tudi

vlogo lastnikov virov podatkov, ki naj bi imeli temeljito znanje o poslovnih procesih. Navedeno znanje je treba vgraditi v postopke ETL, saj omogočajo zagotavljanje višje kakovosti podatkov.

Postopki ETL morajo biti prilagodljivi, saj poslovanje združb ni statično. Poleg navedenega je treba zagotoviti tudi učinkovitost postopkov ETL. Učinkovitost je mogoče povečati, ko postopke testiranja kakovosti podatkov ločimo od dejanskega prenosa podatkov.

4.5.4 Uporaba podatkov

Podatki so lahko popolnoma natančni, vendar če jih uporabnik ne razume, nastopijo težave. Z uporabo podatkov mislimo na različna poročila, poizvedbe, ki jih končni uporabniki uporabljajo v poslovnih procesih.

Temelj ustrezne rešitve je spet meta podatkovni model, ki mora vsebovati informacijo o ustrezni uporabi podatkov. Na žalost večina združb nima ustreznega meta podatkovnega modela. Ko govorimo o uporabi podatkov, mislimo tudi na njihovo dostopnost. Če želimo, da bodo uporabniki uporabljali podatke, morajo biti ustrezno dostopni. Če je ovir za dostop do podatkov preveliko, je velika verjetnost, da uporabniki takšnih podatkov preprosto ne bodo uporabili. Takšno ravnanje lahko ima hude ali celo usodne posledice za posamezne združbe.

5 ČIŠČENJE PODATKOV

Nekakovostne podatke lahko obravnavamo kot bolezen in bolezen lahko zdravimo in preprečujemo. V zvezi s podatki lahko govorimo o »zdravljenju« in »preprečevanju«. Preprečevanje je vedno boljša izbira, o tem ni dvoma. Vendar če »bolezen« nastopi, jo je treba zdraviti in zdravilo za nekakovostne podatke je čiščenje podatkov. Čiščenje po navadi opravljamo z orodji, ki nam močno olajšajo delo. O orodjih lahko govorimo kot o antibiotikih za nekakovostne podatke in njihova uporaba je smotrna v naslednjih primerih (Donohue, Chang in Bostwick 2004):

- prepoznavanje in odprava ponavljajočih se podatkov,
- uveljavitev poslovnih pravil,
- podpiranje obstoječe tehnologije,
- olajšanje dela uporabnikom.

Uporaba različnih orodij za čiščenje podatkov odpira novo vprašanje, in sicer kje čistiti podatke. Podatke lahko čistimo, preden jih zapišemo v skladišče podatkov, vendar s tem ne bomo rešili vseh težav na ravni združbe. Še bolj problematično se zdi čiščenje podatkov na ravni uporabniške rešitve. V tem primeru bi morali čistiti podatke v vsaki uporabniški rešitvi posebej, kar povzroča še višje stroške kot čiščenje podatkov za potrebe skladišča podatkov. Poleg tega spremembe poslovnih pravil otežujejo čiščenje podatkov v vsaki uporabniški rešitvi posebej.
Donohue, Chang in Bostwick (2004) navajajo tudi najpogostejše razloge za uporabo orodij za čiščenje podatkov:

- poenotenje osebnih podatkov v zbirkah podatkov v celotni združbi;
- različni viri podatkov;
- različnost podatkovnih tipov in oblik navajanja podatkov;
- imena so včasih navedena kot posamezen zapis, včasih so razbita na več delov;
- vnos osebnih podatkov najpogosteje poteka prek uporabniškega vmesnika brez deklarativnih omejitev;
- nezaupanje virom podatkov;
- odsotnost mehanizmov uravnavanja kakovosti podatkov;
- odsotnost vrednosti, ki omogočajo enolično določanje zapisov.

Kako se torej lotiti procesa čiščenja podatkov? Dongre (2004) kot glavni cilj procesa čiščenja podatkov navaja ureditev in zapis podatkov v ciljne sisteme, ki lahko poteka popolnoma samodejno, ročno ali pa kot združek samodejnega in ročnega čiščenja. Tierstein (2004) pa kot osrednji cilj čiščenja podatkov navaja odpravljanje napak v podatkih. Podobno opredelitev

ponudita tudi Rahm in Do (2000), ki pravita, da je čiščenje podatkov proces odstranjevanja nepopolnosti v podatkih z namenom povečati kakovost podatkov. Navajata različne težave, s katerimi se moramo spopadati v procesu čiščenja podatkov. Avtorja ločita težave, ki nastajajo v podatkih iz enega vira (težave zaradi nekakovostne podatkovne sheme: kršenje enoličnosti in referenčne integritete; težave pri vnosu podatkov v zbirko podatkov: sintaktične napake in podvajanje), in težave, ki nastajajo v podatkih iz različnih virov (težave, povezane s shemo: nepopolnost; težave zaradi vnosa: nepopolnost).
Ko pregledujemo podatke, iščemo predvsem naslednje nepravilnosti v podatkih:

- izjemne vrednosti,
- manjkajoče vrednosti,
- neusklajene vrednosti in
- nepopolne vrednosti.

5.1 Manjkajoče vrednosti oziroma nepopolne vrednosti

Čiščenje podatkov poskuša zapolniti manjkajoče vrednosti podatkov, odpraviti izjemne ter neusklajene vrednosti. Kako torej zapolnimo manjkajoče vrednosti? Han in Kamber (2001) navajata več metod, in sicer:

- zanemarimo manjkajoče vrednosti – ta metoda je zelo slaba, če je število manjkajočih vrednosti zelo veliko;
- ročno vnašanje manjkajočih vrednosti – takšen način nam lahko vzame veliko časa, zato ni primeren za velike zbirke podatkov z veliko manjkajočimi vrednostmi;
- uporaba stalnice (v nadaljevanju konstante) – vse manjkajoče vrednosti lahko zamenjamo z isto konstanto (na primer »Neznana vrednost«). Ta metoda ni priporočljiva, ker lahko zavede posamezen algoritem odkrivanja vzorcev med podatki;
- uporaba povprečne vrednosti posamezne lastnosti – če, na primer, manjka podatek za plačo, ga lahko nadomestimo s povprečno plačo, ki jo izračunamo iz znanih podatkov o plači;
- uporaba najverjetnejše vrednosti – z različnimi avtomatskimi metodami poskušamo predvideti, katera vrednost je najprimernejša.

Najpogosteje uporabljamo prav zadnjo med navedenimi metodami, metodo uporabe najverjetnejše vrednosti. Pri obravnavi manjkajočih vrednosti se moramo osredotočiti samo na tiste manjkajoče vrednosti, ki v realnem svetu obstajajo, hkrati pa so pomembne za poslovni proces (Müller in Freytag 2003).

PREGLEDNICA 10: PRIMER MANJKAJOČIH VREDNOSTI

A	B	C	Vrednost podatkov					
101	0	120	101	0	0	0	120	0
?	1	140	0	1	1	0	140	0
97	?	200	97	0	0	1	200	0
96	2	?	96	0	2	0	0	1
98	1	160	98	0	1	0	160	0

Vir: prirejeno po Kennedy, Solving data mining problems through pattern recognition, 1997.

V preglednici 10 smo izrecno označili, katera vrednost je prisotna (0) in katera manjka (1). S tem smo opozorili različna orodja in uporabnike na manjkajoče vrednosti ter jim omogočili ustrezno nadaljnje obravnavanje podatkov: na primer SOZP se lahko nauči zanemarjati manjkajoče vrednosti, vendar je potrebna previdnost, saj prevelika količina manjkajočih podatkov lahko povzroči nesmiselne rezultate.

5.2 Izjemne vrednosti

Izjemne vrednosti so napačne vrednosti in vrednosti, ki precej odstopajo od ostalih vrednosti. Najpogosteje uporabljene metode za odpravo izjemnih vrednosti so (Han in Kamber 2001 ; Maletic in Marcus 2000):

- primerjanje, ki določeno vrednost primerja s sosednjimi vrednostmi. Obstaja več različic primerjanja (primerjanje s povprečjem, primerjanje s skrajnimi vrednostmi, primerjanje z mediano);
- razvrščanje – razvrščanje uvrsti vrednosti v posamezne razrede in v vsak razred uvrsti vrednosti, ki so si podobne. Tako bodo v nekem razredu izjemne vrednosti, ki jih potem zanemarimo;
- združek računalniškega in človeškega pregledovanja;
- regresija – izjemne vrednosti lahko odpravimo tudi z regresijo, kjer poskušamo s pomočjo funkcije napovedati, kakšna naj bi bila določena vrednost.

Opravilo čiščenja podatkov lahko izvedemo samodejno z različnimi orodji, ročno ali pa kombiniramo obe metodi.

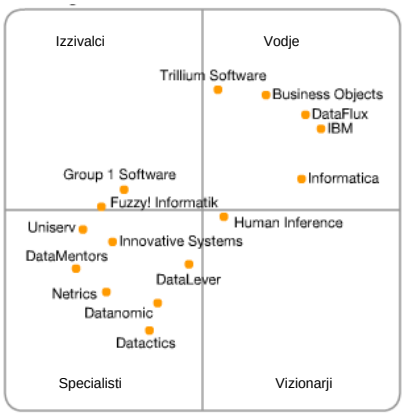
5.3 Samodejno čiščenje podatkov

Samodejno čiščenje se začne s preverjanjem podatkov v zbirkah podatkov. Namen preverjanja je zaznava in opredelitev napak v podatkih, ki je navadno opravljena po obravnavi

funkcionalnosti zbirke podatkov oziroma uporabniške rešitve, ki jo opravi oseba z ustreznim znanjem. Preverjanje naj bi odkrilo kritične napake, tj. napake, ki jih je treba nemudoma odpraviti, nekritične napake, ki nimajo ključnega vpliva na poslovanje združbe ter količino obeh vrst napak. Nova metoda odkrivanja napak v podatkih je SOZP. Osnovna zamisel te metode je uporaba sodobnih algoritmov pri odkrivanju napak v podatkih, na primer uporaba algoritma povezovalnih pravil. Omenjena zamisel izhaja iz dejstva, da je za vsako zbirko podatkov mogoče izluščiti pravila, s katerimi lahko preverjamo kakovost podatkov (Hipp, Guntzer in Nakhaeizadeh 2000). Če je na primer v zbirki podatkov navedena poštna številka 5261, je verjetnost, da je to kraj Šempas, skoraj 100-odstotna. Hipp (2001) s sovtorjema navaja tudi pomembnost obravnavanja pravil, ki imajo manjšo verjetnost. Če je na primer znamka avtomobila Mercedes razred S, bo 90 % teh avtomobilov imelo bencinski motor.

Po opravljenem preverjanju kakovosti podatkov je treba ugotovitve pregledati, kar je spet naloga uporabnikov s poslovnim znanjem. Poročila morajo biti čim bolj pregledna, da omogočajo strokovnjakom temeljit pregled, ki se nadaljuje s samodejnim čiščenjem podatkov. Samodejno čiščenje se po navadi izvaja kot skupek različnih procesov, vendar je treba upoštevati, da v zbirkah podatkov po navadi obstajajo napake, ki jih ni mogoče samodejno odpraviti in njihova odprava zahteva ročni poseg v podatke. Po koncu samodejnega čiščenja podatkov je treba vnovič preveriti, kakšno je stanje podatkov, saj le tako lahko ugotovimo uspešnost procesa čiščenja.

SLIKA 11: GARTNERJEV KVADRANT ORODIJ ZA RAVNANJE S KAKOVOSTJO PODATKOV



Vir: Friedman in Bitterer, Magic quadrant for data quality tools, 2007.

Na trgu obstaja veliko komercialnih orodij za čiščenje podatkov, ki pa imajo po večini isto slabost, tj. da so orodja običajno usmerjena bodisi samo v zaznavanje napak bodisi samo v odpravljanje napak. Zelo malo pa je orodij, ki bi zagotavljala oboje (Raman in Hellerstein 2001). Med drugimi prevladujejo orodja, ki poskušajo odpravljati napake s pomočjo različnih slovarjev (na primer slovar krajev in pripadajočih poštnih številk) oziroma poskušajo povezati entitete v

zbirkah podatkov z realnimi objekti. Razvoj takšnih orodij je zelo drag, hkrati pa imajo orodja omejeno učinkovitost (Luebberes, Grimmer in Jarke 2003). Na sliki 11 je prikazan Gartnerjev kvadrant najboljših orodij za leto 2007, pri čemer pa je treba upoštevati dejstvo, da njihovi analitiki niso popolnoma neodvisni in mnogokrat odločijo oziroma napovejo dogodke v korist največjih svetovnih multinacionalk.

Ker je nabor orodij raznovrsten in njihove lastnosti številne, se management po navadi težko odloči za ustrezno orodje. Goasdouéjeva, Nugier, Duquennoy in Laboissova (2007) so pripravili spisek pomembnih lastnosti za posamezno področje ukvarjanja s kakovostjo podatkov.

5.4 Ročno čiščenje podatkov

Ročno čiščenje podatkov je potrebno pri napakah v podatkih, ki jih samodejno čiščenje ne more odpraviti. Tudi ta proces se začne s preverjanjem napak v podatkih, katerega izhod je poročilo o napakah v podatkih. Po odkritju napak se začne ročno odpravljanje napak, ki ga izvajajo strokovnjaki s posameznih področij. Po odpravi napak je treba izvesti revizijo, da se ugotovi uspešnost ročnega odpravljanja napak.

Stroški, pa tudi napor in potreben čas, za ročno odpravljanje napak so neprimerno višji v primerjavi s samodejnim čiščenjem napak. Posledično je treba pozorno ugotoviti vzroke za pojav takšnih napak in jih poskušati tudi odpraviti, saj vnovični pojav in potreba po ročnem odpravljanju napak pomenita ogromne stroške za posamezno združbo.

5.5 Združek čiščenja podatkov

Večina napak v podatkih ima značilnosti z obeh navedenih področij, torej s področja samodejne odprave napak in s področja ročnega popravljanja podatkov. Značilnosti napak narekujejo vrsto čiščenja podatkov, a na končno odločitev za vrsto čiščenja podatkov odgovornih oseb v združbi vplivajo še številni drugi dejavniki.

Na izbor samodejnega čiščenja podatkov naj bi vplivali naslednji dejavniki:

- količina podatkov – če imamo ogromno napak v podatkih, je samodejno čiščenje podatkov veliko bolj primerno kot ročno čiščenje podatkov;
- vse ali večino napak lahko odpravimo samodejno s pomočjo logičnih pravil;
- stroški ročne odprave napak z vidika časa so neprimerno višji od stroškov samodejne odprave napak;
- zamenjava SSOP in potreba po pripravi podatkov za nov SSOP.

Ročno odpravljanje napak pa je smotrno v naslednjih primerih:

- napak ni mogoče odpraviti samodejno s pomočjo logičnih pravil;
- majhna količina podatkov, ki jo je treba popraviti, zaradi česar je ročni način ekonomsko smotrnejši.

Združek čiščenja podatkov pa se uporablja, kadar so napake v podatkih enakomerno razporejene med tiste, ki jih lahko odpravimo samodejno, in tiste, ki jih moramo odpraviti ročno. Pri procesu čiščenja podatkov je pomembno spremljanje učinkovitosti navedenega procesa z vidika izboljšanja kakovosti podatkov, pri čemer je izboljšanje kakovosti treba spremljati za vsako skupino napak posebej.

6 MERJENJE KAKOVOSTI PODATKOV – PREREZ PODATKOV

Prerez podatkov je opredeljen kot skupek analitičnih tehnik za ugotavljanje strukture, vsebine in kakovosti podatkov v različnih strukturah podatkov oziroma zbirkah podatkov (Olson 2003). Marshall (2007) opredeljuje prerez podatkov kot analitično metodo, ki se uporablja pri pregledu zbirke podatkov z namenom ugotoviti probleme, povezane s kakovostjo podatkov, medtem ko Miriyala (2007) o prerezu podatkov govori kot o proučevanju podatkov z namenom ugotoviti razlike med dejansko in pričakovano vrednostjo ter obliko zapisa.

Rezultat procesa prereza podatkov je ocena kakovosti podatkov v zbirki podatkov, ki vsebuje poročilo o popolnosti podatkov, o težavah podatkov (razvrščenih po pomembnosti) ter o porazdelitvi navedenih težav v zbirki podatkov. Metoda prereza podatkov je drugačna od tradicionalnih analitičnih metod, saj slednje ne omogočajo ekonomsko upravičenih analiz v sodobnih zbirkah podatkov. Metoda prereza podatkov izboljšuje tradicionalne analitične metode na treh področjih:

- avtomatizacija procesa ugotavljanja kakovosti podatkov – čas analize se lahko zmanjša tudi do 90 %;
- analiziranje velike količine podatkov, saj se ni treba omejiti samo na vzorec podatkov iz zbirke podatkov, temveč lahko analiziramo celotno zbirko podatkov;
- obravnavanje zapletenih pravil med podatki v zbirki podatkov.

V procesu ugotavljanja kakovosti podatkov z metodo prereza podatkov sodeluje manjše število strokovnjakov – analitikov, ki jim podporo ponujajo drugi strokovnjaki, po navadi poznavalci poslovnih procesov v združbi. Druga zelo koristna skupina ljudi, ki po potrebi pomaga analitikom, prihaja iz vrst informatikov, ki so zgradili uporabniško rešitev ali pa jo vzdržujejo. Pri tem se predpostavlja, da informatiki dobro poznajo podatkovne strukture, ki so predmet metode prereza podatkov.

Vhod v navedeni proces so podatki in meta podatki, ki dejansko opredeljujejo kakovost podatkov in so nepogrešljivi z vidika ugotavljanja kakovosti. Olson (2003, 124) ugotavlja, da bi imeli pri popolnem meta podatkovnem modelu zelo lahko delo pri iskanju težav v podatkih. Vendar večina podatkovnih modelov sploh nima opredeljenega meta podatkovnega modela oziroma je slednji zelo pomanjkljiv. Olson (2003) navaja tudi najpomembnejše sestavine meta podatkovnega modela, in sicer:

- podatki o virih podatkov, zbirki podatkov ter načinu dostopa do zbirke podatkov;

- podatki o lastnostih posameznega stolpca: glavni ključi, tuji ključi, sprožilci, shranjeni postopki (lastnosti so odvisne od vrste zbirke podatkov);
- podatki o najpomembnejših pravilih, ki spreminjajo podatke.

Zbrane podatke je treba preveriti in obravnavati z vidika različnih strokovnjakov, saj le tako lahko zagotovimo večjo kakovost meta podatkov in posledično večjo kakovost metode prereza podatkov.

Metoda prereza podatkov se uporablja za objektivno ocenjevanje kakovosti podatkov in je dopolnilo subjektivnemu ocenjevanju kakovosti podatkov, ko uporabniki podatkov izrazijo svoje videnje kakovosti podatkov. V raziskovalnem delu knjige se uporablja subjektivno ocenjevanje kakovosti podatkov, vendar sem želel s predstavitvijo metode prereza podatkov pokazati tudi drugo možnost, ki jo imajo posamezne združbe.

6.1 Cilji prereza podatkov

Glavni cilj metode prereza podatkov je ugotovitev kakovosti podatkov v zbirki podatkov, ki jo lahko opišemo z meta podatki ter razlikami med pričakovano in dejansko vrednostjo podatkov. Meta podatki so tako vhod kot izhod procesa prereza podatkov. Ko govorimo o izhodu, mislimo predvsem na podatke oziroma zapise, ki jih je proces preverjanja zavrnil. Takšni zapisi morajo vsebovati najmanj:

- naziv preglednice in stolpca;
- oznako vrstice, kjer se je zgodila kršitev;
- pravilo, ki je bilo kršeno;
- napačne vrednosti s svojimi frekvencami;
- celotno število vrstic, ki so bile preizkušene.

Na tem mestu je treba poudariti previdnost pri izdelavi različnih matrik kakovosti podatkov. Različno oblikovanje poročil nas lahko hitro zavede, na primer število napačnih podatkov z različnih vidikov, število kršitev poslovnih pravil ipd. Oblikovanje poročil oziroma matrike kakovosti podatkov ima veliko pozitivnih lastnosti. Morda je najpomembnejša koristnost matrike dokaz managementu, da je proces preverjanja kakovosti podatkov učinkovit in prinaša otipljive dokaze, da so »podatki v težavah«. Drugi pomemben pozitivni vidik izdelave matrike je možnost merjenja učinkovitosti ukrepov za izboljšanje kakovosti podatkov. Ukrepi so po navadi povezani s stroški in management mora videti napredek v izboljšavi kakovosti podatkov. Z metodo prereza ugotovimo kakovost podatkov pred uvedbo ukrepov in po uvedbi ukrepov ter rezultate med seboj primerjamo. Izvedba preverjanja kakovosti podatkov združbe nima vedno enakega stroškovnega učinka, saj meta podatke zberemo le enkrat, pozneje popravljamo po

potrebi, zato je z omenjeno metodo najdražje prvo preverjanje kakovosti podatkov. Pri primerjavi dveh matrik moramo biti previdni, da se merila za njihovo izdelavo niso pomembno spremenila. Analitik je lahko v času od izdelave ene do izdelave druge matrike pridobil nova spoznanja o poslovanju, o podatkih, ki jih lahko vgradi v novo matriko, in s tem delno ali popolnoma onemogoča primerjavo skozi čas (Bauer 2004).

Po pregledu pozitivnih lastnosti izdelave matrike kakovosti podatkov je treba pregledati tudi negativne lastnosti. Matrike pokažejo samo na dejstvo, da težave obstajajo, medtem ko odpravljanja težav neposredno ne omogočajo. Kljub vloženemu trudu in uporabi sodobnih metod je velika verjetnost, da ne bomo uspeli prepoznati vseh napak v zbirki podatkov. Pri tej trditvi smo lahko še natančnejši. Četudi uporabimo metodo razreza podatkov in nam omenjena metoda potrdi, da v naši zbirki podatkov ni napak, obstaja velika verjetnost, da napake v zbirki podatkov dejansko obstajajo. Zato lahko trdimo, da nam uporaba te metode omogoča odkriti napake v podatkih, vendar pokazati število napak ali morebitno odsotnost napak presega njene zmožnosti. Z izdelavo matrike se delo v združbi ne konča. Prelaganje odgovornosti za sprejemanje ukrepov, ki naj izboljšajo kakovost podatkov, na sodelavce, je navadno neučinkovito in lahkomiselno dejanje. Takšne naloge pogosto ostanejo nizko na prednostni lestvici ljudi, ki so bili zanje zadolženi.

Ko govorimo o ugotavljanju kakovosti podatkov v zbirki podatkov, mislimo predvsem na razsežnost natančnost ter delno na razsežnosti popolnost in doslednost. Logično je, da z omenjeno metodo ne moremo proučevati razsežnosti zaupanja v podatke, ki je ena od bistvenih razsežnosti matrike kakovosti podatkov, vendar slednja bolj spada na področje psihologije.

6.2 Model prereza podatkov

Metoda prereza podatkov vsebuje različne dejavnosti oziroma tehnike, ki jih lahko uporabimo pri odkrivanju kakovosti podatkov. Te dejavnosti oziroma tehnike so:

- odkrivanje (angl. *discover*),
- preverjanje trditev,
- vidni pregled podatkov,
- preverjanje meta podatkov.

Dejavnosti niso neodvisne, ampak se med seboj prepletajo. Z odkrivanjem preverjamo dejstva o podatkih, ki so analitiku znana oziroma o njih domneva. V podatkih analitik lahko išče določene vzorce (na primer pojavljanje posebnih znakov : , . =) in je ob ponavljanju vzorca pri velikem številu podatkov blizu potrditvi svojih domnev. V tem pogledu se zdi tehnika odkrivanja celo močnejša kot tehnika preverjanja trditev. To lahko potrdimo s primerom odkrivanja vrednosti v

stolpcu, ki naj bi vseboval vrednosti med 0 in 10.000. Odkrivanje nam pokaže, da je večina vrednosti med 200 in 300, kar je opozorilo analitiku, da je nekaj narobe s podatki oziroma da se stolpec najbrž ne uporablja za prvotni namen.

Preverjanje trditev pokaže analitiku, ali podatki trditvi ustrezajo ali ne, in če ne, kje so odstopanja. To tehniko po navadi izvedemo skupaj s tehniko odkrivanja. Enostaven primer preverjanja trditve je trditev, da vsi podatki vsebujejo vrednosti med 0 in 10.000. Za preverjanje trditve lahko pripravimo poizvedbo, ki nam vrne podatke, kjer trditev ne drži. Takšna poizvedba bi lahko bila: *Select #Izdelka, NazivIzdelka From Artikli Where Cena not between 0 and 10 000*.

Vidni pregled podatkov je pomembna metoda pri odkrivanju nepravilnosti v podatkih, saj nam lahko nakaže težave, ki jih s prej omenjenima metodama ne utegnemo zaznati. Vidni pregled po navadi vsebuje:

- frekvenčno porazdelitev vrednosti v stolpcu;
- primerjavo seštevkov iz različnih virov podatkov;
- podatke, ki odstopajo od neke normalne vrednosti.

Pri vidnem pregledu po navadi uporabljamo različne pripomočke, ki nam dodatno olajšajo delo. Različni slikovni prikazi, uporaba preglednic s podatki, ki so ustrezno urejeni, lahko močno olajšajo pregled podatkov.

Poleg omenjenih pristopov je potrebno nenehno preverjanje meta podatkov, saj le tako lahko zagotovimo kakovostne podatke skozi čas.

6.3 Tehnike prereza podatkov

Uporaba metode prereza podatkov ima svoje značilnosti. Podatke začnemo presojati v skladu z metodo »od spodaj navzgor«. Značilnost takšne obravnave podatkov je v postopnem odkrivanju in odpravljanju napak. Prednost takšne metode je predvsem v natančnosti oziroma zanesljivosti odkrivanja napak v primerjavi z nasprotno metodo, tj. »od zgoraj navzdol«. Presoja podatkov poteka v naslednjih korakih:

- analiza stolpca,
- strukturna analiza,
- analiza povezav med podatki.

Z analizo stolpcev poskušamo poiskati morebitne napačne vrednosti, s strukturno analizo ciljamo na odkrivanje napačnih združkov posameznih pravilnih vrednosti, kar je tudi naloga tretjega koraka, tj. analize povezav med podatki. S slednjo pa poskušamo naše védenje o

podatkih še nekoliko obogatiti, in sicer poskušamo odkriti »čudne« vrednosti podatkov. Kljub podrobni analizi podatkov bo v zbirki podatkov ostal delež nekakovostnih podatkov (razsežnost natančnost), ki jih z metodo prereza podatkov ne bomo mogli odkriti.

Analiza stolpcev nam torej omogoča presojo vrednosti podatkov v posameznem stolpcu neodvisno od drugih stolpcev. Primer: polje #poslovalnice na izdanem računu določene združbe lahko obsega vrednosti od 001 do 010 in vsakršna vrednost, ki ne spada v omenjeni obseg, je označena kot napačna vrednost. Odkrivanje napak z analizo stolpca je odvisno predvsem od poznavanja lastnosti posameznega stolpca. To ne pomeni samo poznavanja fizičnih lastnosti, temveč tudi vsebinskih pravil. Odsotnost znanja o vsebinskih lastnostih posameznega polja pomeni možnost spregleda sintaktično pravilnih, vendar semantično napačnih podatkov.

Strukturna analiza podaja znanje o medsebojni odvisnosti posameznih stolpcev in medsebojni odvisnosti posameznih preglednic. Torej so predmet obravnave strukturne analize glavni ključi, tuji ključi, morebitni odvečni stolpci in druge referenčne omejitve. Sodobni SUBP-i omogočajo učinkovito uvedbo omenjenih značilnosti, medtem ko podatki v starejših datotečnih sistemih pogosto kršijo omenjena pravila. Uporaba strukturne analize je zato še posebno pomembna pri prenosu podatkov iz datotečnih sistemov v relacijske zbirke podatkov, saj bo ustrezen FIRPM zavrnil uvoz podatkov, ki kršijo omenjena pravila. Strukturna analiza sicer ugotovi, katera množica podatkov je nekakovostna, vendar ni sposobna odkriti posameznega nekakovostnega podatka znotraj množice podatkov.

Zadnji korak pri metodi razreza podatkov je obravnavanje povezav med podatki (podatkovna pravila). Podatkovna pravila so podmnožica poslovnih pravil. Za lažje razumevanje si poglejmo naslednji primer. Pravilo, ki pravi, da je uporaba luksuznih službenih vozil dovoljena le za delovno mesto trgovskega potnika, je podatkovno pravilo, medtem ko je pravilo, ki pravi, da se stranki iz skupine A pri prodaji obračuna 10-odstotni popust, poslovno pravilo. Toda »surova« podatkovna pravila po navadi niso primerna za preverjanje podatkov, zato jih je treba pretvoriti v ustrezno obliko, ki omogoča samodejno preverjanje.

Poglejmo si posamezne korake bolj podrobno.

6.3.1 Analiza stolpcev

Stolpec je osnovni gradnik preglednic in pregled kakovosti podatkov se mora začeti v stolpcih. Pri pregledu posameznih stolpcev poskušamo ugotoviti, katere in kakšne so tiste lastnosti določenega stolpca, ki opredeljujejo sprejemljive vrednosti, poleg tega pa poskušamo odkriti tudi vrednosti, ki te lastnosti kršijo (napačne vrednosti). Tipične pomembne lastnosti posameznega stolpca so:

- ime stolpca,
- vsebinski pomen (opis),
- podatkovni tip,
- dolžina,
- dovoljene vrednosti,
- null/not null,
- edinstvenost.

Ime stolpca je izredno pomembno, saj naj bi nakazovalo vsebino, ki jo stolpec hrani. Sodobni sistemi omogočajo dovolj dolga imena stolpcev, da jih lahko smiselno poimenujemo, medtem ko so starejše uporabniške rešitve omogočale le zelo kratka imena. Ime stolpca naj bo torej takšno, da bo uporabnikom podatkov dovolj hitro jasno, kakšna je vsebina posameznega stolpca. V preglednici, kamor se bodo shranjevali podatki o kontih za knjiženje prispevkov za izobraževanje zasebnika, je naziv stolpca »KontoZaKnjizenjePrispevkaZaIzobrazevanje« veliko boljši kot naziv »Konto1«.

Sorodna imenu stolpca je tudi lastnost »vsebinski pomen«, ki obširneje določa vsebino posameznega stolpca. Če se vrnemo k prejšnjemu primeru, bi v opisu zapisali, da je stolpec »KontoZaKnjizenjePrispevkaZaIzobrazevanje« polje, kamor se shranjujejo vrednosti kontov iz razreda 4, ki so namenjeni knjiženju temeljnice za potrebe prispevka za izobraževanje posameznega zaposlenega.

S podatkovnim tipom natančneje določimo vrednosti, ki jih posamezen stolpec lahko sprejme, z lastnostjo »dolžina« pa pravila za vnos vrednosti še dodatno zaostrimo. Dolžina stolpca nam lahko zelo koristi pri pregledu podatkov, saj v primeru, ko ima stolpec opredeljeno dolžino 30, v njem pa je največ dvomestnih znakov, sklepamo, da je s podatki nekaj narobe. Vnos vrednosti v posamezen stolpec lahko omejimo tudi z deklarativnimi omejitvami.

O vrednostih null/not null je bilo že veliko napisanega pri opredeljevanju matrike kakovosti podatkov, zato o tej lastnosti ne bomo več podrobneje pisali. Edinstvenost je lastnost posameznega stolpca, ki določa, da se vrednosti v stolpcu ne smejo ponavljati. S tega vidika je edinstvenost podobna glavnemu ključu, vendar je med njima tudi pomembna razlika. Edinstvenost, ki jo navadno določimo z edinstvenim indeksom, omogoča vrednosti null, medtem ko jih glavni ključ ne omogoča.

S pregledom navedenih lastnosti posameznega stolpca uspemo ugotoviti nekakovostne podatke. Takšen pregled, če je izveden »ročno«, je zelo zamuden in neučinkovit, zato je pri velikem številu stolpcev in vrednosti stolpcev treba uporabiti ustrezno programsko rešitev, ki nam močno olajša delo.

6.3.2 Strukturna analiza podatkov

Po pregledu stolpcev se je treba lotiti pregleda struktur, ki so opredeljene oziroma bi morale biti opredeljene med podatki z namenom ugotoviti morebitne težave. Pri strukturni analizi se moramo osredotočiti na medsebojne odvisnosti, tj. funkcijske odvisnosti med stolpci ter iskanje sinonimov, tj. stolpcev, ki imajo enak pomen. Strukturna analiza mora potekati znotraj posamezne preglednice ter med preglednicami v zbirki podatkov.

Ko govorimo o funkcijskih odvisnostih, govorimo o normalizaciji, o kateri je bilo že ogromno govora, zato o njej na tem mestu ne bomo več govorili. Analiza funkcijskih odvisnosti nam pokaže glavne ključe, tuje ključe, denormalizirane tabele ter izvedene stolpce.

Pri sinonimih imamo več možnosti, in sicer razmerje glavni ključ v eni preglednici in pripadajoč tuji ključ v drugi preglednici. V preglednici »Delavec« imamo glavni ključ opredeljen na stolpcu »DelavecID« in stolpec z enakim imenom (lahko ima tudi drugačno ime) je tudi v preglednici »Placa«, kjer je tuji ključ. Sinonimi lahko nakazujejo odvečnost določenega stolpca, s čimer se spet dotikamo normalizacije. Polje »Opis« v preglednici »Artikli« se lahko pojavi tudi v preglednici »GlavaRacuna«. Poleg navedenih sinonimov se lahko srečamo s sinonimi, ki nimajo nobene vsebinske povezave.

Strukturna analiza nam lahko odkrije tudi pravila, ki so pomembna pri ravnanju s podatki, na primer pri prenosu podatkov v skladišče podatkov iz različnih virov podatkov. Neizvedba tega koraka lahko povzroči velike težave v procesu prenosa podatkov. Prav lahko se zgodi, da denormaliziran vir podatkov in denormaliziran ciljni sistem povzročita nekakovostne podatke, tj. nenatančne vrednosti pri različnih agregiranjih. Poleg tega so strukturna pravila v virih podatkov mnogokrat kršena, ker fizično niso vzpostavljena, in prenos v ciljni sistem, ki ima strukturo vzpostavljeno, propade oziroma ga je treba vnovič izvesti. Posledica vnovične izvedbe prenosa so dodatni stroški in časovne zamude.

Ko govorimo o strukturni analizi, govorimo tudi o izdelavi idealnega podatkovnega modela. Najboljši podatkovni model naj bi bil model v tretji normalni formi, vendar v praksi le redko najdemo takšne modele, saj so pri načrtovanju podatkovnega modela velikokrat pomembni kompromisi. Ti kompromisi izhajajo predvsem iz nekdanje slabše zmogljivosti programske in strojne opreme, ko so načrtovalci počasnejše delovanje poskušali zaobiti z denormalizacijo. Danes za kaj podobnega ni razlogov, razen morda zaradi specifičnih poslovnih pravil. Primerjava med dejanskim in najboljšim podatkovnim modelom nam da osnovo za tretji korak v metodi prereza podatkov, tj. analizo povezav med podatki.

6.3.3 Analiza povezav med podatki

Analiza povezav pomeni preveriti podatkovna pravila med podatki, ki so podmnožica poslovnih pravil. Pravila nam omogočajo statičen pogled na podatke, torej v danem trenutku. Podatkovna pravila sicer ločimo na enostavna in zapletena, vendar je ne glede na vrsto pravila cilj analize povezav med podatki isti: odkriti pravila. Rezultat je besedni opis posameznega pravila, ki pa ga moramo nadgraditi tudi z logičnim opisom. Poleg delitve pravil na enostavna in zapletena obstaja še delitev pravil na dosledna in nedosledna. Za dosledna pravila velja, da morajo podatki tem pravilom popolnoma ustrezati, da jih lahko opredelimo kot natančne, na primer datum rojstva nikoli ne more biti mlajši od datuma zaposlitve. Druga vrsta pravil pa so nedosledna pravila, kjer posamezna kršitev še ne pomeni nujno nenatančnega podatka.

Pravila navadno prepoznamo s pomočjo različne listin, če obstajajo. Naslednja možnost za prepoznavanje pravil je pregled izvorne kode, kjer si lahko pomagamo z različnimi programskimi rešitvami. Pregled poslovnih procesov ter uporaba zdravega razuma nam prav tako lahko odkrijeta pomembna podatkovna pravila.

Analiza povezav med podatki lahko odkrije mnogo pravil, ki sploh niso upoštevana v naši uporabniški rešitvi. Takšna pravila je treba po posvetu z ustreznimi strokovnjaki, ki so navadno projektanti uporabniških rešitev, vgraditi v uporabniške rešitve in lahko pomenijo dodatna preverjanja pri vnosu podatkov ter pri obdelavi podatkov.

7 RAVNANJE S KAKOVOSTJO PODATKOV

Če smo o čiščenju podatkov govorili kot o antibiotikih, torej zdravljenju, je zdaj čas, da si pogledamo tudi preprečevanje, torej procese, ki zagotavljajo boljšo kakovost podatkov. Ti procesi so sestavni del ravnanja s kakovostjo podatkov. Kakovost podatkov je treba skrbno načrtovati (Redman 2001).

Ravnanje s kakovostjo podatkov zajema določanje vlog uporabnikov podatkov ter postopkov pri zbiranju, shranjevanju, vzdrževanju ter uporabi podatkov (Geiger 2004), pri čemer je odločilno sodelovanje managementa in strokovnjakov s področja informatike. Management je zadolžen za določitev poslovnih pravil, ki zadevajo kakovost podatkov, hkrati pa prevzema objektivno odgovornost za kakovostne podatke. Strokovnjaki s področja informatike pa so zadolženi za tehnični vidik, tj. zbirke podatkov, strežnike, mrežno infrastrukturo ipd.

Uvedba ravnanja s kakovostjo podatkov pa ni preprosta naloga, saj se mora zainteresirana združba soočiti s številnimi težavami (Redman 2001):

- pogosto se noben oddelek ne čuti odgovornega za nekakovostne podatke oziroma je za nekakovost podatkov obdolžen informacijski oddelek;
- ravnanje s kakovostjo podatkov zahteva sodelovanje med različnimi poslovnimi funkcijami. V združbi je lažje vzpostaviti navpično sodelovanje v primerjavi z vodoravnim sodelovanjem, saj je med različnimi oddelki pogosto ljubosumje. Vendar sodobne oblike združb spodbujajo prav takšno naravnanost, tj. procesno ureditev;
- navadno mora združba priznati, da ima številne težave s kakovostjo podatkov. Za takšno dejanje je po navadi potreben znaten dogodek, ki zadolžene v združbi opozori na težave. Če takšnega dogodka ni, management združbe najverjetneje ne bo trošil denarja za nekaj, kar po njegovem mnenju ne obstaja;
- uvedba ravnanja s kakovostjo podatkov zahteva disciplino med zaposlenimi. Disciplina pomeni, da se morajo zaposleni natančno držati pravil, ki jih opredeli management združbe, na primer dosledno vnašanje podatkov o strankah;
- izvajanje povzroča stroške in zahteva dejavnosti zaposlenih;
- proces ravnanja s kakovostjo podatkov je delovno intenziven in trajen proces;
- koristi kakovostnih podatkov je številčno težko opredeliti, še posebno ker se zaposleni po navadi zavedajo nekakovosti, zato so našli različne možnosti, kako zaobiti morebitne težave.

Težave pri ravnanju s kakovostjo podatkov še natančneje opredeli Redman (2004), ki navaja 12 mogočih ovir. Pri tem izpostavi dve, ki se po njegovem mnenju pojavljata v večini združb, in sicer

nerazumevanje oziroma slabo razumevanje povezav med kakovostjo podatkov in poslovanjem združbe ter napačna porazdelitev zadolžitev med zaposlenimi. V zvezi s kakovostjo podatkov in poslovanjem združbe je zanimiv Abilenov paradoks, ki pravi, da četudi se večina zaposlenih zaveda pomembnosti kakovosti podatkov za poslovanje združbe, bo združba kot celota kakovost še vedno zanemarjala (Harvey 1988).

Uvedba ravnanja s kakovostjo podatkov zahteva nenehno izobraževanje zaposlenih, ki morajo spoznati, da so kakovostni podatki pomembni za uspešno poslovanje združbe. Poleg izobraževanja je treba nujno vzpostaviti sodelovanje med različnimi oddelki ter zagotoviti tehnično podporo. Poudariti je treba, da je vzrok za uvedbo ravnanja s kakovostjo podatkov največkrat odzivanje na nastale težave, zelo redko pa so združbe pri ukvarjanju s kakovostjo podatkov proaktivne (Loshin 2004). Za proaktivno metodo je treba zgraditi ustrezno matriko kakovosti podatkov ter analizirati morebitne težave. Za kaj takega je potrebna ustrezna motivacija managementa, ki pa iz zgoraj omenjenih razlogov le redko poseže po proaktivnem ravnanju s kakovostjo podatkov.

Za ravnanje s kakovostjo podatkov potrebujemo ustrezno metodo in TIQM je ena od njih.

7.1 TIQM

TIQM je celostna metoda ravnanja s kakovostjo podatkov (Eppler in Helfert 2004), ki omogoča opredeliti stroške zaradi nekakovostnih podatkov ter koristi izboljšanja kakovosti. Prvotno je bila poznana kot TDQM (*Total Data Quality Management*). Preimenovanje v TIQM je nastalo kot posledica opažanja, da management veliko bolje zaznava priporočila o kakovosti, ko je govor o kakovosti informacij namesto podatkov. Management namreč šteje podatek predvsem za tehnični izraz. TIQM združuje spoznanja različnih avtorjev z različnih področij, in sicer Demingova spoznanja o kakovosti podatkov (14 točk), Masaakijevo metodo KAIZEN, Juranovo metodo načrtovanja kakovosti in Crosbijeva spoznanja (2007).

TIQM ni programska rešitev, temveč je filozofija, tako kot obvladovanje razmerij s kupci. Pomemben del filozofije so tudi priporočila in smernice za preoblikovanje združbe iz navpične v vodoravno združbo, kar omogoča doseganje boljše kakovosti. Ravnanje s kakovostjo zahteva ustrezne listine in izobraževanje, kar TIQM omogoča.

Uvajanje TIQM sestoji iz šestih korakov, ki pa niso nujno zaporedni, ampak lahko začne združba izvajati kateri koli korak:

- opredelitev podatkov in potrebne infrastrukture,
- opredelitev kakovosti podatkov in postopkov merjenja kakovosti,

- merjenje stroškov in posledic nekakovostnih podatkov,
- popravljanje oziroma čiščenje podatkov,
- izboljšanje procesov pridobivanja podatkov,
- vzpostavitev poslovnega okolja, ki zagotavlja kakovostne podatke oziroma informacije.

Razumevanje kakovosti podatkov je pravzaprav poslovni, in ne sistemski problem, saj bistvo TIQM ni v izboljšanju kakovosti podatkov v zbirkah podatkov, temveč izboljšanje poslovne učinkovitosti z zmanjšanjem količine nekakovostnih podatkov. Za dosego tega cilja se je treba osredotočiti predvsem na ponudnike in uporabnike podatkov, in ne samo na podatke. To izhaja iz spoznanj metode KAIZEN, ki je sestavni del TIQM. Po metodi KAIZEN je treba kakovost izboljšati pri izvoru, in ne na koncu vrednostne verige (Imai 1986). Izvor je pri tej metodi poimenovan kot GEMBA (Imai 1997) in pri podatkih je GEMBA mesto, kjer se podatki pridobivajo ali izmenjujejo.

Pri obravnavanju kakovosti se ne smemo osredotočiti samo na vidik ustreznosti, ampak tudi na vidik natančnosti in predstavitve, ki sta pomembna vidika matrike kakovosti podatkov, ju pa veliko združb zanemarja. To še posebno velja za ponudnike programskih rešitev za presojo kakovosti podatkov (English 2002). Pomembna je tudi Englisheva ugotovitev, da bo poslovanje združb, ki ne bodo sprejele programa za zagotavljanje kakovosti podatkov, v prihodnosti ogroženo.

Del TIQM je tudi obravnavanje informacije kot poslovnega učinka ter izdelava informacijske mape, kar si bomo pogledali v nadaljevanju.

7.1.1 Informacija kot poslovni učinek

Ko govorimo o informaciji kot poslovnem učinku vlečemo vzporednice s proizvajalnim sistemom. Osnovna zamisel je v možnosti prodajanja informacije različnim porabnikom. S takšno metodo lahko opredelimo štiri različne vloge (Wang 1998):

- dobavitelji informacij – zbiralci podatkov za proces proizvajanja informacij;
- proizvajalci informacij – izvajajo proces pretvarjanja podatkov v informacije;
- odjemalci informacij – uporabniki, ki uporabljajo informacije pri svojem delu;
- skrbniki procesa proizvajanja informacij – odgovorni so za celoten proces ustvarjanja informacij.

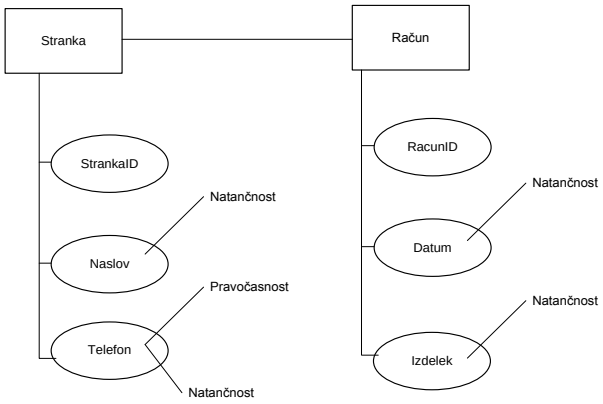
Opredelitev informacije kot produkta ima dva pomembna izhoda:

- entitetno-relacijski diagram, ki opredeljuje informacijo ter njen kakovostni vidik;

- orisan proizvajalni sistem, ki pretvarja podatke v informacije in prikazuje povezave med ponudniki, uporabniki, proizvajalci informacij ter nadzorniki proizvajalnega procesa.

Na sliki 12 je prikazan primer razširjenega entitetno-relacijskega diagrama, kjer so izpostavljene razsežnosti kakovosti podatkov, ki so pomembne za našo uporabniško rešitev. Primeri informacije kot poslovnega učinka so na primer rojstni list, račun v trgovini, mesečni bančni izpisek. Za ravnanje z informacijo kot poslovnim učinkom je pomembno, da ima združba mehanizem za vzpostavitev informacijske mape (Wang et al. 2003), katere izdelavo si bomo pogledali v nadaljevanju.

SLIKA 12: RAZŠIRJENI ENTITETNO-RELACIJSKI DIAGRAM Z VIDIKOM KAKOVOSTI



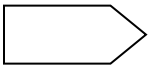
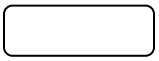
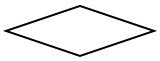
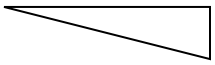
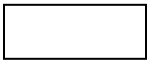
7.1.2 Informacijska mapa

Informacijsko mapo izdelamo s pomočjo osmih različnih simbolov, ki so predstavljeni v preglednici 11. Omenjeni simboli omogočajo izdelavo informacijske mape, ki managerjem omogoča jasen pogled na »proizvajanje« informacij s poudarkom na kritične korake, ki najbolj vplivajo na kakovost.

Poleg tega informacijska mapa omogoča prepoznavanje ozkih grl pri proizvodnji informacij, lastništva procesov, meje združbe in informacijskega sistema ter merjenje kakovosti v različnih korakih proizvajanja informacij (Shankaranarayanan, Wang in Ziad 2000).

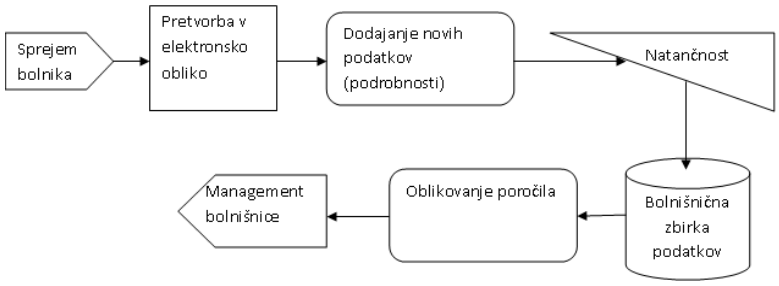
Izdelavo informacijske mape je najlažje prikazati na praktičnem primeru. V nadaljevanju sledi primer delovanja dela bolnišnice (sprejem bolnika), povzet po Shankaranarayananu, Wangu in Ziadu (2000). Informacijski proizvod procesa sprejema bolnika v bolnišnico je dnevno, tedensko in mesečno poročilo o bolnikih, ki zajema število bolnikov ter povprečni čas ležanja v bolnišnici, kar omogoča načrtovanje zasedenosti posameznih oddelkov.

PREGLEDNICA 11: SIMBOLI INFORMACIJSKE MAPE

Simbol	Pomen
	Začetek – vhod surovih podatkov
	Proces – kakršno koli ravnanje z vhodnimi podatki
	Shranjevanje podatkov – datotečni sistemi in/ali zbirke podatkov
	Odločitev
	Kakovost – preverjanje kakovosti podatkov
	Meje informacijskega sistema – meja, ki določa, kako podatki prehajajo iz enega informacijskega sistema v drug informacijski sistem
	Meje združbe – meja, kjer podatki prehajajo iz enega poslovnega okolja(lahko tudi oddelka) v drugo poslovno okolje (oddelek)
	Konec – uporabnik informacij

Vir: Wang et al., An information product approach for total information awareness, 2003, str. 7, preglednica 2.

SLIKA 13: PRIMER IZDELAVE INFORMACIJSKE MAPE



Vir: prirejeno po Shankaranarayanan et al., IP-MAP: representing the manufacture of an information product, 2000.

Vhod v proizvajanje informacij v procesu sprejema bolnika v bolnišnico so podatki, ki izvirajo od bolnika, ter podatki, ki jih dajo zaposleni v bolnišnici.

K diagramu je treba dodati še meta podatkovni model, ki opredeljuje posamezne gradnike.

PREGLEDNICA 12: META PODATKOVNI MODEL ZA INFORMACIJSKO MAPO

Naziv	Oddelek	Lokacija	Proces	Sestavljen iz	Osnovni sistem
Sprejem	Sprejemna pisarna	Urgenca	Ustaljen postopek		Pacientov zdravstveni karton

Vir: prirejeno po Shankaranarayanan et al., IP-MAP: representing the manufacture of an information product, 2000.

Z informacijsko mapo lahko uveljavljamo načelo odgovornosti, kar omogoča uvedbo kakovosti v vsakem koraku proizvodnje informacij. Poleg tega pa je zelo koristna tudi v načrtovanju proizvajalnega časa posamezne informacije.

TIQM lahko izvedemo v okviru metode *Six Sigma*, zato si v nadaljevanju poglejmo, kaj je *Six Sigma in kakšno vlogo ima*.

7.2 Standardizacija in problemi s standardi

Zagotavljanje kakovosti podatkov je mogoče tudi z vpeljavo različnih standardov. Trenutno obstaja že nekaj poskusov postavitve standardov, povezanih predvsem z osebnimi imeni in naslovi, a na žalost je razširjenost uporabe omenjenih standardov zelo omejena. Razlog za majhno razširjenost je zelo preprost. Osebna imena in naslovi so preveč zapleteni, da bi jih lahko umestili v okvir standardov, čeprav njihova uporaba prinaša kar nekaj prednosti, predvsem pri izmenjavi podatkov. Vendar so omejitve na žalost prevelike, in dokler bodo obstajale, standardi ne bodo dosegli večje veljave. Pogoje za razširjenost standardov navaja Rhind (2002), ki pravi, da morajo biti standardi brezplačni, splošno sprejeti, njihovo sprejemanje oziroma uvedba pa mora biti preprosta.

Če upoštevamo Rhindove (2002) utemeljitve ter jih primerjamo z razpoložljivimi standardi za imena in naslove, lahko najdemo kar nekaj kršitev njegovih utemeljitev.

Zelo očitno je, da obstaja veliko število različnih standardov, med katerimi je nekaj izključujočih, zato je splošno sprejemanje posameznega standarda skoraj nemogoče, saj se uporabnik ne more odločiti, katerega naj uporabi. Za uporabnike bi bilo lažje, če bi imeli manjše število lažje dostopnih in dobro opisanih standardov, saj so mnogokrat prav slabi meta podatki krivi za neuvedbo določenega standarda.

Težave lahko nastopijo tudi zaradi prevelikega navdušenja nad hitro uvedbo določenega standarda, še posebno če določen standard šele razvijajo. V tem primeru je uvedba skoraj zagotovo obsojena na propad, saj združba hitro ugotovi, da od spreminjajočih se standardov ne

more pričakovati ustreznih učinkov, še posebno če so spremembe med posameznimi različicami velike. Poleg omenjene težave so ovira tudi pričakovanja managementa v združbi. Po začetnem navdušenju se navadno izkaže, da standard pokriva le del problemskega področja, kar onemogoča doseganje prvotno postavljenega cilja, ki je pogosto nestvaren, saj predvideva popolno rešitev težav z osebnimi imeni in naslovi.

Pri splošnem sprejemanju standardov je treba izpostaviti še eno zelo očitno oviro, tj. kulturno raznolikost. Kulturna raznolikost preprosto ne omogoča sprejetja enotnega standarda za osebna imena ter naslove. Obvladovanje osebnih imen in naslovov je zelo težavno že znotraj posamezne države, v mednarodnem prostoru pa se težave ustrezno potencirajo. Rhind (2002) tudi ugotavlja, da se združbe zaradi stroškov, časa in drugih potrebnih prvin, ki so potrebne za uvedbo ter sledenje določenega standarda, raje odločajo, da standardov ne bodo upoštevale.

7.3 Ravnanje s kakovostjo podatkov s stališča revizije

Ravnanju s kakovostjo namenja v zadnjem času posebno pozornost združba ISACA, ki združuje strokovnjake s področja informatike in katere namen je svetovanje na različnih področjih IT, na primer ravnanje z informatiko, ravnanje z informacijsko varnostjo ter vzpostavitev vzvodov nadzorovanja procesov informatike (2007).

Pri ravnanju s kakovostjo podatkov ISACA svetuje šest ključnih vprašanj, na katere mora vsaka združba poiskati odgovore (Adler 2006):

- Ali v imamo vzpostavljen proces ravnanja s kakovostjo podatkov?
  - Kdo je odgovoren za ravnanje?
- Kakšno metodo uporabljamo za ocenjevanje obstoječega stanja?
  - Ali obstajajo kakšne primerljive vrednosti?
- Kakšna je naša strategija?
  - Kako bomo odpravili razkorak med obstoječim in želenim stanjem?
- Koliko so vredni naši podatki?
  - Koliko prihodkov prinašajo podatki?
  - Kakšni so stroški nekakovostnih podatkov?
- Kje so naše slabosti?
  - Kako izračunamo tveganje?
- Kako merimo naš napredek?
  - Kaj nam meritve povedo?

S stališča kakovosti podatkov je predvsem pomembno izračunati vrednost podatkov v posamezni združbi, kjer je jasno stališče, da je vrednost podatkov odvisna od vrednosti IT. Adler

(2006) pri ocenjevanju vrednosti podatkov zavzema tržno načelo, torej je vrednost podatkov odvisna od vsote, ki so jo uporabniki pripravljeni plačati, zato tudi predlaga, da je najboljša pot za oceno vrednosti podatkov zgraditev notranjega trga v združbi, kjer se srečajo ponudniki in povpraševalci po posameznih podatkih. Takšne metode pa ni mogoče vedno uporabiti, vsaj pri etično-moralnih vprašanjih, kjer je vrednost podatka neprecenljiva in je ne moremo opredeliti s preprosto tržno metodo.

ISACA poudarja tudi pomembnost meta podatkovnega slovarja pri ravnanju s kakovostjo podatkov, pri čemer model izstopa predvsem kot pomoč skrbnikom kakovosti na vseh treh ravneh: izvedbeni, taktični in strateški ravni (Livesley 2006). Livesley (2006) v svojem delu predstavlja združbo *BMO Financial Group* in poudarja vlogo meta podatkovnega slovarja pri poslovanju svoje združbe, zato je s stališča nepristranskega opazovalca tematika malce subjektivno naravnana. Vsekakor pa ne smemo spregledati vloge meta podatkovnega slovarja, saj omogoča opredelitev ene resnice oziroma enotnega pogleda na pomembne entitete. S tem se izognemo zmešnjavi, ko imajo različni oddelki drugačen pogled na iste podatke ter iste poslovne procese. Vendar za enoten pogled na poslovne procese meta podatkovni slovar ni dovolj, saj si morajo združbe v tem primeru pomagati tudi s procesom opredeljevanja poslovnih procesov, ki je izjemno pomemben za enoten pogled na poslovne procese (Groznik in Vičič 2007).

Ravnanje s kakovostjo podatkov je izredno pomembno področje, ki se intenzivno razvija. Ravnanje je usmerjeno predvsem v preprečevanje nekakovostnih podatkov (preventivo), čeprav odpravljanje napak (kurativa) zavzema precejšen delež njegovih dejavnosti.

8 POMEMBNEJŠI SKLEPI IZ TEORETIČNIH SPOZNANJ

Po proučevanju svetovne literature je mogoče postaviti kar nekaj sklepov, ki so usmerjali raziskavo. V nadaljevanju sledijo naslednji sklepi:

- ločevanje izrazov »podatek« in »informacija«,
- pomembnost kakovosti podatkov za posamezno združbo,
- razširjena opredelitev kakovosti podatkov,
- opredelitev kakovosti izvedbe relacijskega podatkovnega modela.

8.1 Sklep 1: Ločevanje izrazov podatek in informacija

Za potrebe razrešitve raziskovalnega problema je treba razčistiti razmerje podatek – informacija. Čeprav se v odnos med podatkom in informacijo pogosto vključi še pojem sporočilo[13] (Mihelčič 1972, 24), slednjega v knjigi ne upoštevam. In čeprav je v teoriji semiotike ločnica med podatkom ter informacijo jasna, te ločnice v knjigi namenoma ne upoštevam. Pri prebiranju literature s področja kakovosti podatkov je bilo mogoče opaziti, da avtorji pri opredeljevanju kakovosti ne ločijo med podatki in informacijami, ampak izraza uporabljajo izmenjaje oziroma jim je podatek sinonim za informacijo in nasprotno (Wang 1998). Takšna metoda je uporabljena tudi v knjigi. Vendar je treba kljub temu navesti razlike med podatkom in informacijo, kot izhaja iz teoretičnih izhodišč v literaturi, čeprav se ta opredelitev v knjigi ne uporablja.

Za podatke lahko ugotovimo, da so nizi znakov, ki pojasnjujejo pojave. So torej opredmetenje oziroma predstavitev dejstev, pojmov, predstav in znanja. Podatek je na neki način ugotovljeno in zapisano dejstvo, ki izraža neko stanje oziroma dogodek v sistemu ali njegovem okolju. Pravzaprav se v njih izražajo dogodki in stanja stvarnega ali miselnega sveta. Informacija ima določene lastnosti, ki jih podatek nima. Eno od gledanj oziroma opredeljevanj informacij izhaja iz semiotike, vede o znakih in znakovnih sistemih ter njihovih splošnih značilnostih. Semiotika trdi, da lahko govorimo o informaciji šele, ko so izpolnjene tri njene osnovne razsežnosti: sintaktična, semantična in pragmatična.

Sintaktična razsežnost opredeljuje formalno pravilnost zapisa v določenem znakovnem sistemu. Nanaša se torej le na izbrani znakovni sistem, ki pa mora biti takšen, da je z njim mogoče izraziti želeno sliko, pojav. Sintaktična razsežnost torej obsega celoto povezav med znaki v določenem znakovnem sistemu. Sintaksa danega znakovnega sistema določa, kako v množici znakov oblikujemo in preoblikujemo formalno pravilna sporočila, na primer v slovenščini iz besed

[13] Sporočilo lahko opredelimo kot vsebino komuniciranja, komuniciranje pa je oddajanje in sprejemanje sporočil (Mihelčič 2003, 387)

pravilne stavke, oblika zapisa datuma, uporaba rdeče barve na semaforju kot signala za ustavitev vožnje. Semantična razsežnost izraža vsebino formalno pravilnega sporočila, in to na osnovi zveze z objektom realnega ali pa miselnega sveta. Semantična razsežnost se torej ukvarja s celoto povezav med znaki in vsebino, ki jo znaki in sporočila nosijo, o tem, kar označujejo. Pogoj za semantično razsežnost je sintaktično pravilno sporočilo. Pragmatična razsežnost je dosežena takrat, ko na osnovi osmišljenega sporočila pade odločitev, katere učinek je ukrepanje, delovanje oziroma akcija. Pragmatična razsežnost sporočila je uresničljiva šele v povezavi s subjektom – odločevalcem. Ta edini lahko presodi, v kolikšni meri je sporočilo uporabno v konkretnem odločitvenem procesu. Pragmatična razsežnost se torej ukvarja s celoto povezav med znaki in subjekti, z vrednostjo znakov in sporočil (za odločevalca) v danih odločitvenih razmerah. Pogoj za doseganje pragmatične razsežnosti sporočila in s tem informacije je sintaktično pravilno in semantično doseženo sporočilo (Lesjak et al. 2006).

Šele ko so izpolnjene vse tri razsežnosti, ki se kažejo v smiselnem delovanju, ukrepih, smemo govoriti o informaciji. Potemtakem mora biti informacija v danem znakovnem sistemu sintaktično pravilna, imeti mora nedvoumno vsebino, za odločevalca mora imeti pragmatično (uporabno) vrednost.

Podatki so torej niz znakov, ki opisujejo neko dejstvo, pojav, objekt, osebo, pri čemer morajo biti sintaktično in semantično pravilni. Čeprav nekateri dosledno zagovarjajo ločitev terminov podatek in informacija, pa iz novejše literature izhaja težnja po zmanjšanju razlik med pojmovanjem omenjenih izrazov. Vzrok za takšno postopanje očitno izhaja iz nemoči pri doseganju enotnosti pri opredeljevanju izrazov podatek in informacija. Poleg tega je mogoče sklepati, da je ločevanje na izraza podatek in informacija za neukega uporabnika prezahtevna ter odvečna naloga, še posebno če je končni uporabnik management, ki navadno nima časa posvečati energijo teoretičnemu ločevanju. V tem primeru se izraz podatek spremeni v informacijo, ki je managementu združb bolj poznan.

Zelo razširjen že omenjeni vidik ločevanja na izraza podatek in informacija je vidik semiotike. Če bi dosledno spoštovali smernice semiotike, bi pri opredeljevanju razsežnosti kakovosti podatkov lahko kmalu zašli v težave. Z vidika semiotike je namreč informacija podatek, ki ima sintaktično, semantično in pragmatično vrednost. Na tem mestu se pojavi vprašanje, ali ima podatek, katerega razsežnosti pravočasnost in dostopnost sta nekakovostni, tj. podatek je nedosegljiv ali je dosegljiv nepravočasno, sploh možnost, da ima oziroma hrani pragmatično vrednost za končnega uporabnika ter s tem postane informacija. Odgovor je dokaj preprost: ne. Podatek, ki ga z vidika nameravane uporabe ne dobimo pravočasno oziroma je nedostopen, nam ne koristi, torej z vidika semiotike ne more postati informacija. Podobno je tudi z razsežnostma ustreznost in razumljivost. Podatek, ki ni ustrezen ali pa ga uporabnik ne razume, ne more imeti

pragmatične vrednosti, zato ne more nikoli postati informacija. Še slabše je z razsežnostjo natančnost, saj podatek, ki ni natančen za nameravano uporabo, sploh ne dosega meril sintaktične pravilnosti in semantične vrednosti.

Torej, če bi v knjigi upoštevali navedeno delitev na podatke in informacije, bi naleteli na veliko težav z opredeljevanjem razsežnosti kakovosti informacij, ki opredeljujejo njeno uporabno vrednost, v nekaterih primerih pa celo sintaktični in semantični vidik. Nesmiselno bi bilo govoriti o natančnosti, pravočasnosti, ustreznosti, dostopnosti, razumljivosti kot o razsežnostih kakovosti informacij, saj so slednje smiselne le v povezavi s pojmom podatek. Informacija, ki ni pravočasno dostopna, ustrezna in razumljiva, nima nobene uporabne vrednosti za odločevalca, torej sploh ne more biti informacija, ampak je lahko samo podatek.

Da bi se izognili morebitnim težavam pri ukvarjanju s kakovostjo podatkov, se pojma podatek in informacija v knjigi razumeta kot sopomenki in se uporabljata izmenjaje. S tem sledimo sodobnim smernicam na področju opredeljevanja in ocenjevanja kakovosti podatkov.

8.2 Sklep 2: Pomembnost kakovosti podatkov za posamezno združbo

Podatki igrajo temeljno vlogo v vsakem informacijskem sistemu, tj. v SSOP ter sistemih za poslovno obveščanje, hkrati pa so jedro filozofije obvladovanja razmerij s kupci. Če povzamemo Herrensteinov izrek, da je vsako vedenje odločanje, za vsako odločanje pa potrebujemo podatke, potem so podatki dejansko eden od temeljev vsake združbe. Naše odločanje bo lahko boljše le s kakovostnejšimi podatki. Kakovostni podatki so jedro analitičnih sistemov, ki so že zdaj zelo pomembni, njihove možnosti pa so lahko še precej večje.

V analitičnih sistemih izstopa predvsem sistem SOZP. Združbe so namreč ves čas prisiljene iskati tekmovalne prednosti, ki bi jim povečale možnosti za zaslužek ter posledično preživetje na trgu. Pozornost združb se je od blaga oziroma storitev ter z njimi povezanih lastnosti (cene, kakovosti, dobavnih rokov) preusmerila na posameznega kupca. Kaj pomaga združbi zelo kakovosten poslovni učinek po ugodnih cenah, če ne zadovoljuje potreb posameznega kupca?! Vendar pa je kupec danes veliko bolj zahteven kot nekoč, lahko bi rekli, da je postal zelo razvajen. Sodobna IT ter razvoj občil omogočata opravljanje nakupov iz naslanjača, opravljanje storitev prek svetovnega spleta in še bi lahko naštevali. Ves razvoj v tehnologiji omogoča posamezniku, da lahko lažje in bolje izrazi svoje potrebe, združba mora le poslušati. Idealno bi bilo, da bi združba zaznala posameznikove potrebe, še preden bi jih posameznik izrazil. SOZP omogoča združbam spoznati svojega kupca, njegove navade. To ni nič novega, včasih je vez med kupcem in prodajalcem že obstajala. Pred velikim razmahom potrošniške družbe so ljudje navadno hodili kupovat blago oziroma storitve ves čas k istemu trgovcu. Slednji je kmalu poznal vsakega posameznika, predvsem v smislu, kaj posamezen kupec rad kupi ter katere poslovne učinke po

navadi kupi skupaj. Ko sta se trgovec in kupec bolje spoznala, je trgovec že lahko predvideval kupčeve potrebe ter temu prilagodil svojo ponudbo. Toda z rastjo trga in prehodom v potrošniško družbo se je stik med trgovcem in kupcem pretrgal. Kmalu je bil posamezen kupec le še anonimen kupec, ki je kupoval v kakšni veliki trgovski verigi.

Združbe, predvsem trženjski oddelki so se v zadnjih letih začeli zavedati, da morajo kupca spet spoznati, zato so začele zbirati njegove podatke. Pri vsakem stiku med kupcem in združbo imata oba možnosti spoznavati drug drugega. Če sta včasih stike navezovala bolj redko, z obiskom kupca v prodajalni ali na sejmu ter z obiskom trgovca na domu, je danes zaradi razvoja IT komuniciranje bolj pogosto. Največji vpliv ima danes internet, ki omogoča obiskovanje in komuniciranje 24 ur na dan, vse dni v letu. Kupec na ta način lahko izrazi svoje potrebe, želje pa tudi pripombe. Nič več mu ni treba nemo sprejemati vsega, kar mu trgovec ponudi, in če se trgovec ne odzove na njegove pripombe, lahko odide k tekmecem. Prav ta odhajanja porabnikov k tekmecem so začele združbe izredno skrbeti, ker je zelo drago dobiti novega kupca, še dražje pa spet dobiti starega kupca. Združbe poskušajo danes kupca obdržati toliko časa, dokler koristi od kupca presegajo stroške, da ga ohranimo zadovoljnega, torej dokler stroški ohranjanja kupca ne presegajo dobička od kupca.

Z najnovejšo tehnologijo kupci spet prihajajo iz anonimnosti. Za ta korak pa združbe potrebujejo kakovostne podatke. Mercatorjeva kartica Pika, Tuševa kartica niso nič drugega kot plačilo izdajatelja kartic porabnikom v zameno za njihove osnovne demografske podatke in tudi njihove podatke o nakupovalnih navadah. Vsakič ko imetnik kartice svojo kartico dejansko tudi uporabi, se v zbirko podatkov zapišejo njegovi nakupi. Naslednji korak izdajatelja kartic je analiza podatkov ter sprejemanje primernih odločitev.

Primer neizkoriščenega potenciala podatkov so tudi finančne združbe, ki z oblikovanjem profila imetnika različnih kartic lahko preprečijo marsikatero nepooblaščeno transakcijo in si prihranijo stroške. A za kaj takega v slovenskem prostoru očitno še ni zanimanja. Bo že držala trditev, da v združbah sicer vlada prepričanje o pomembnosti podatkov za njihovo preživetje, vendar s podatki ne ravnajo ustrezno.

Kljub vsemu povedanemu moramo veliko pozornost posvetiti zasebnosti posameznika, ki je v zadnjem času izredno zapostavljena. Analize, ki jih opravljamo, morajo imeti privoljenje posameznika, poleg tega moramo spoštovati tudi zakone, družbene in moralne.

8.3 Sklep 3: Subjektivno in objektivno merjenje kakovosti podatkov

V literaturi je že opisanih nekaj različnih metod, ki so prilagojene ožjim problemskim področjem (Huang, Lee in Wang 1999 ; Laudon 1986), medtem ko metoda, ki bi veljala v večini primerov,

manjka. Pipino, Yang in Wang (2002) predlagajo združevanje subjektivnega in objektivnega ocenjevanja pri določanju kakovosti podatkov.

Subjektivna ocena kakovosti je močno zaznamovana z znanjem in potrebami udeležencev v procesu ravnanja s podatki in njihovo dojemanje kakovosti podatkov močno vpliva na njihovo ravnanje. Subjektivno ocenjevanje je navadno izvedeno z vprašalnikom, v katerem anketiranci izrazijo svoje videnje kakovosti podatkov. V raziskovalnem delu je uporabljena prav takšna metoda za ocenjevanje kakovosti podatkov v PIS.

Objektivno ocenjevanje lahko izvedemo z upoštevanjem poslovnih pravil in omejitev, organizacijskih in državnih predpisov ali pa navedenih dejavnikov ne upoštevamo. Pri objektivnem ocenjevanju kot mero kakovosti podatkov Pipino, Lee in Wang (2002) predlagajo uporabo preprostega razmerja, agregatni funkciji min in max ter metodo uteženega povprečja.

Metoda razmerja

V statistiki se pri določanju oziroma potrjevanju trditve pogosteje uporablja metoda, s katero poskušamo dokazati, da trditev ne drži (če tega ne moremo dokazati, potrdimo pravilnost odločitve), kot pa metoda, s katero poskušamo dokazati, da trditev drži. Enako metodo poskušamo uporabiti, ko ocenjujemo kakovost podatkov, torej ocenjujemo nekakovost podatkov namesto kakovosti podatkov: $K_p = 1 - \frac{število\ nekakovostnih\ podatkov}{število\ vseh\ odatkov}$. Metodo razmerja je smiselno uporabiti pri obravnavanju popolnosti, doslednosti in natančnosti. Za vsako razsežnost moramo postaviti natančne opredelitve in postopek obravnave v izogib subjektivnemu ravnanju udeležencev, na primer zamenjan vrstni red črk v besedi je v določenemu primeru dovoljen in ga dovoljujemo, v drugem primeru pa ni dovoljen.

Agregatni funkciji min in max

Agregatni funkciji sta primerni za obravnavo bolj zapletenih razsežnosti, in sicer zaupanja, pravočasnosti, količine podatkov in dostopnosti. Funkcija min pomeni previdnejši pristop, kjer najslabša ocena, ki jo dobimo na primer z metodo razmerij, prevlada za oceno razsežnosti, medtem ko funkcija max pomeni optimistični pristop k ocenjevanju razsežnosti. V slednjem primeru prevlada najboljša ocena razsežnosti. Agregatna funkcija min se navadno uporablja za oceno zaupanja v podatke in ocenjevanje ustreznosti količine podatkov, funkcija max pa se uporablja v primeru pravočasnosti in dostopa. Uporabo navedenih agregatnih funkcij je najlažje pokazati na praktičnem primeru. Vzemimo primer, v katerem zaupanje podatkov ocenjujemo na podlagi verodostojnosti virov podatkov ter predhodnih izkušenj, pri čemer prva spremenljivka

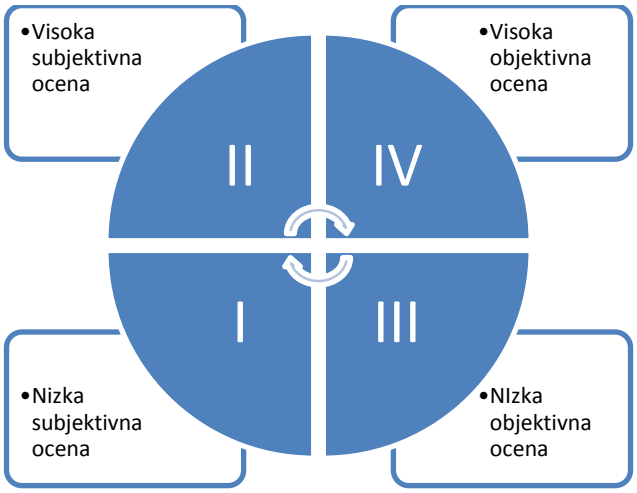
dobi oceno 0,8, druga pa 0,6. Ker uporabljamo funkcijo min, je ocena razsežnosti zaupanje enaka slabši vrednosti med obema ocenama, torej 0,6.

Uteženo povprečje

Uteženo povprečje lahko v določenih primerih uporabimo kot nadomestilo za agregatni funkciji min in max. Če uporabimo zgornji primer, bi verodostojnosti virov podatkov pripisali utež 0,7, predhodnim izkušnjam pa utež 0,3. Skupna ocena kakovosti zaupanja bi bila v tem primeru 0,74.

Po izvedbi subjektivne in objektivne ocene kakovosti podatkov je treba obe oceni primerjati in razrešiti morebitna odstopanja. Za primerjanje rezultatov subjektivne in objektivne ocene lahko uporabimo matriko, ki jo predlagajo Pipino, Lee in Wang (2002).

SLIKA 14: MATRIKA PRIMERJAVE SUBJEKTIVNE IN OBJEKTIVNE OCENE



Vir: *Ballou et al., Modeling information manufactoring systems to determine information quality product, 1998.*

Cilj analize kakovosti podatkov je umestitev rezultatov v posamezno rubriko, pri čemer je najboljša novice za združbo, če analiza zapolni rubriki II in IV, medtem ko rubriki I in III zahtevata nadaljnjo obravnavo (slika 14). Obravnava je potrebna tudi pri zelo odstopajočih rezultatih analiz, na primer hkratna zapolnitev rubrik II in III. Loshin (2005) pri izdelavi matrike kakovosti podatkov poudarja stališča, ki izhajajo iz vsakega dobrega projekta, torej jasnost, merljivost, ustreznost glede na poslovanje, možnost nadzora, kakovost predstavitve, kakovost poročanja, sledljivost in zmožnost prikazovanja podrobnosti (angl. *drill down*). V svojem delu se dotakne tudi kritike sodobnih orodij, ki kakovost podatkov prikazujejo brez povezave s poslovanjem združbe. Loshin (2005) meni, da bi vsaka matrika oziroma meritev kakovosti podatkov morala upoštevati vpliv na učinkovitost poslovanja, dobiček/izgubo, tveganje in tudi nemerljive oziroma težje razumljive dejavnike poslovanja.

Vendar pa je tudi z upoštevanjem Loshinovih (2005) priporočil navedena analiza še vedno statična, saj kakovost podatkov ne meri z vidika nameravane uporabe in z vidika procesa ravnanja s podatki.

Temeljita ocena kakovosti podatkov, še posebno če je to prva ocena, je zelo pomembna, saj združbi zagotavlja osnovo za spremljanje različnih dejavnosti in njihov vpliv na kakovost podatkov v prihodnosti, in sicer izboljšanje ali poslabšanje kakovosti podatkov. Nancy Mullen (2003) priporoča naslednje korake pri ocenjevanju kakovosti podatkov:

- v združbi je treba izvesti raziskavo med zaposlenimi oziroma uporabniki, katere razsežnosti kakovosti podatkov so najpomembnejše in na katerih področjih v združbi je največ težav s kakovostjo podatkov;
- z znanjem o problemskih področjih lahko opredelimo ustrezne dejavnosti, navadno v obliki projektov, katerih cilj je izboljšanje kakovosti podatkov na najbolj občutljivih področjih v združbi;
- dejavnosti je treba proučiti z vidika stroškov in koristi ter za izbrane dejavnosti določiti roke za izvedbo;
- ponavljanje predhodnih dejavnosti v časovnem razmiku 12–18 mesecev.

Vse navedene analize kakovosti podatkov so dolgotrajne in zahtevajo sodelovanje mnogo ljudi ter strokovnjakov. Zaradi dolgotrajnosti in zahtevnosti omenjenih analiz Fuchs (2002) predlaga bolj praktično metodo, ki je primerna tudi za datotečne sisteme, razpredelnice ter zbirke podatkov in lahko poda hitro oceno kakovosti podatkov. Fuchs predlaga naslednje korake:

- Opredeliti je treba polja, ki bodo uporabljena za poslovno obveščanje in vsako polje tudi analizirati. Pri analiziranju posameznega polja si najlažje pomagamo z razvrščanjem vrednosti, na primer od najmanjše do največje. Rezultat razvrščanja so vrednosti, v katerih lahko hitro prepoznamo manjkajoče in čudne vrednosti.
- Nadaljujemo z analizo povezanih polj, na primer država in oznaka države, stranka in šifra stranke. Z omenjeno analizo ugotovimo, ali so povezani podatki kakovostni ali ne.
- Naslednji korak je preverjanje referenčne integritete, na primer ali imamo vse šifre strank na izdanih računih tudi v šifrantu strank oziroma ali so vse države, iz katerih so naši kupci, vnesene v šifrant držav. Pravilno uporabljanje transakcij, uporaba tujih ključev ter ustrezna raven osamitve transakcije preprečujejo nepravilne vrednosti.
- V zbirkah podatkov oziroma splošno v podatkih je treba preveriti tudi poslovna pravila oziroma poslovno logiko. Poslovna logika je navadno opredeljena v uporabniškem vmesniku ali v kakšnem drugem sloju uporabniške rešitve. Pomembno je, da čim več poslovnih pravil opredelimo v zbirki podatkov, še posebno glavna poslovna pravila.

- Po oceni kakovosti podatkov je treba nekakovostne podatke popraviti.

Olson (2003) k Fuchsovim priporočilom dodaja še postopek pregledovanja zbirnih podatkov v določenem časovnem obdobju, na primer prihodek od prodaje v določeni združbi niha med 1000 in 1500 enotami. Če analiza zbirnih podatkov pokaže odstopanja od povprečja, je velika verjetnost, da določeni podatki manjkajo. Drugo dopolnilo je pregledovanje vrednosti določenih stolpcev, za katere ne moremo postaviti jasno določenega pravila in je potrebno široko znanje o poslovanju združbe. Olson (2003) navedena priporočila imenuje metoda analize, ki je preprostejša in hitrejša kot metoda vnovičnega preverjanja. Slednja zahteva od analitika, da vrednost podatka primerja z izvirno vrednostjo, česar pa analitik ne more vedno storiti. Priporočila, ki jih omenjata Fuchs (2002) in Olson (2003), so dobrodošla za hitro oceno kakovosti podatkov, vendar združbe potrebujejo dolgoročnejše metode oziroma rešitve, ki so bolj zahtevne. Če ugotovimo, da so naši podatki zelo slabi, moramo ukrepati. Še boljše bi bilo, če slabih podatkov sploh ne bi bilo, kar pa zahteva načrtovanje kakovosti podatkov že v samih temeljih, tj. v načrtovanju podatkovnega modela. Dokazovanje vpliva kakovosti podatkovnega modela na kakovost podatkov je naloga knjige in z njo povezane izkustvene raziskave, vendar je za uspešno raziskavo in ugotavljanje povezave treba opredeliti razsežnosti kakovosti relacijske podatkovne sheme oziroma FIRPM.

8.4 Sklep 4: Opredelitev kakovosti podatkov

Po proučitvi literature se torej poraja želja po združitvi različnih metod ter po poenotenju opredelitve kakovosti podatkov, ki bi zajela različne metode in zmanjšala nedoslednosti vsake v knjigi omenjene teorije. Celostna opredelitev izraza kakovost podatkov zajema naslednja spoznanja:

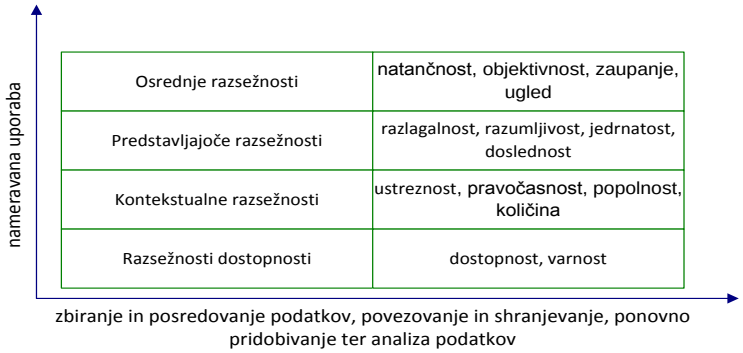
- tradicionalno opredelitev podatkov, povzeto po Wangu in Strongovi,
- Olsonovo opredelitev kakovosti podatkov z vidika nameravane uporabe,
- Dasujevo in Johnsonovo dinamično opredelitev.

Za merjenje kakovosti podatkov uvedemo spremenljivko Q. Enačba (1), ki hrani celotno oceno kakovosti podatkov v vseh korakih ravnanja s podatki, je naslednja:

$$Q = f(K),\ K = (K_1, K_2, ..., K_n),\ 0 \le K_i \le 1, 0 \le Q \le 1. \quad (1)$$

K_i je kakovost podatkov v posameznem koraku ravnanja s podatki, kjer *i* pomeni naslednje korake: zbiranje in posredovanje, povezovanje, shranjevanje in analiza. Ocena celotne kakovosti podatkov leži v obsegu, katerega spodnjo in zgornjo vrednost lahko opredelimo. Spodnjo vrednost obsega dobimo s predpostavko, da so nekakovostni podatki pri procesu zbiranja drugi kot pri povezovanju.

SLIKA 15: OPREDELITEV KAKOVOSTI PODATKOV S POVEZAVO STATIČNE IN DINAMIČNE OPREDELITVE



Spodnjo mejo tako določimo kot zmnožek ocen kakovosti v posameznem koraku $K_1 \cdot K_2 \cdot \ldots \cdot K_n$. Zgornjo vrednost pa dobimo ob predpostavki, da so nekakovostni podatki zmeraj iz iste množice podatkov, zato je zgornja meja določena kot min(K_n). Povedano drugače:

$$Q \in [K_1 \cdot K_2 \cdot \ldots \cdot K_n, \min(K_i)].$$

Naslednji korak je izdelava matematične enačbe za opredelitev kakovosti podatkov v posameznem koraku kot povezava ocene posamezne razsežnosti z namenom uporabe. S spremenljivko D_j izrazimo oceno kakovosti posamezne razsežnosti. D_j lahko zavzame vrednost v razmiku 0 in 1, kjer je 1 najboljša, 0 pa najslabša vrednost. Namen uporabe opišemo s spremenljivko U_j, s katero oceno razsežnosti popravimo navzdol, če menimo, da so podatki za nameravano uporabo slabši, kot je bilo ugotovljeno z merjenjem posamezne razsežnosti. In nasprotno, če menimo, da so podatki boljši, lahko z vrednostjo U_j oceno kakovosti izboljšamo. Učinek nameravane uporabe lahko tudi nevtraliziramo, in sicer z vrednostjo $U_j = 0$. Kakovost podatkov v posameznem koraku je zdaj odvisna od vrednosti posameznih razsežnosti ter namena uporabe, kar lahko zapišemo tudi v naslednji obliki:

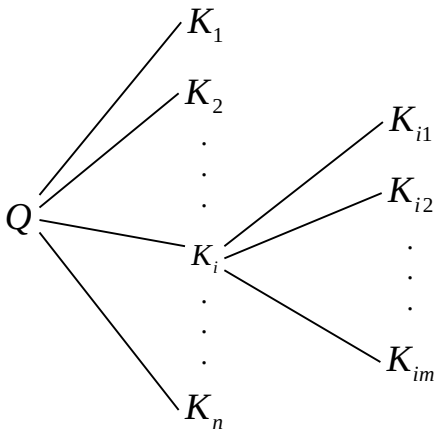
$$K_{ij} = g(D_j, U_j). \tag{2}$$

Zgoraj omenjeno funkcijo lahko zapišemo kot:

$$K_{ij} = g(D_j, U_j) = \begin{cases} Dj+(1-Dj)\cdot Uj, 0 \le Uj \le 1 \\ Dj\cdot(1+Uj), -1 \le Uj < 0 \end{cases}. \tag{3}$$

Spremenljivka U_j ima različno vrednost pri posamezni razsežnosti D_j, saj ima posamezna razsežnost po navadi različno vrednost za določeno problemsko stanje.

SLIKA 16: SHEMA OPREDELITVE KAKOVOSTI PODATKOV



Enačbo, s katero izražamo celotno kakovost ter kakovost v posameznem koraku, je treba normalizirati, da bo vrnjena vrednost vedno med 0 in 1. S tem zagotovimo doslednost pri merjenju kakovosti ter se izognemo morebitnim nejasnostim. Ključni spremenljivki *Q* in *K* tako lahko zavzameta katero koli vrednost med 0 in 1, bližje sta 1, boljša je ocena podatkov. In nasprotno, bližje sta 0, slabša je ocena kakovosti podatkov. Ta zahteva se vidi tudi v enačbi (1).

Tudi kakovost podatkov v posameznem koraku (K_i) lahko ocenimo kot razmik z najmanjšo in največjo vrednostjo. Najmanjša, pesimistična ocena je spet zmnožek ocen posameznih razsežnosti, ki smo jih predhodno že popravili navzgor, navzdol ali pa pustili nespremenjene. Spodnja meja je torej določena kot $K_{i1} \cdot K_{i2} \cdot \ldots \cdot K_{im}$, zgornja meja pa je določena kot $\min_j (K_{ij})$.

Enačba za kakovost podatkov v posameznem koraku tako lahko zavzame vrednosti v obsegu:

$$K_i \in [K_{i1} \cdot K_{i2} \cdot \ldots \cdot K_{im}, \min_j (K_{ij})].$$

Enačbo lahko uporabljamo na različne načine, na primer brez upoštevanja namena uporabe ($U_j = 0$), za objektivno ali subjektivno merjenje kakovosti podatkov. S takšno metodo lažje odkrijemo vzroke za nekakovost podatkov, saj nam analiza pokaže, v katerem koraku procesa ravnanja s podatki se pojavi slabša kakovost. Ta se po navadi brez posredovanja prenese tudi v nadaljnje korake.

Kakovost podatkov je torej mera podatkov, ki nam pove, koliko so podatki natančni, objektivni, vredni zaupanja, ugledni, razlagalni, razumljivi, jedrnati, dosledni, ustrezni, pravočasni, popolni, količinsko primerni, dostopni in varni za nameravano uporabo v procesu zbiranja in posredovanja podatkov, povezovanja in shranjevanja, vnovičnega pridobivanja ter analize podatkov.

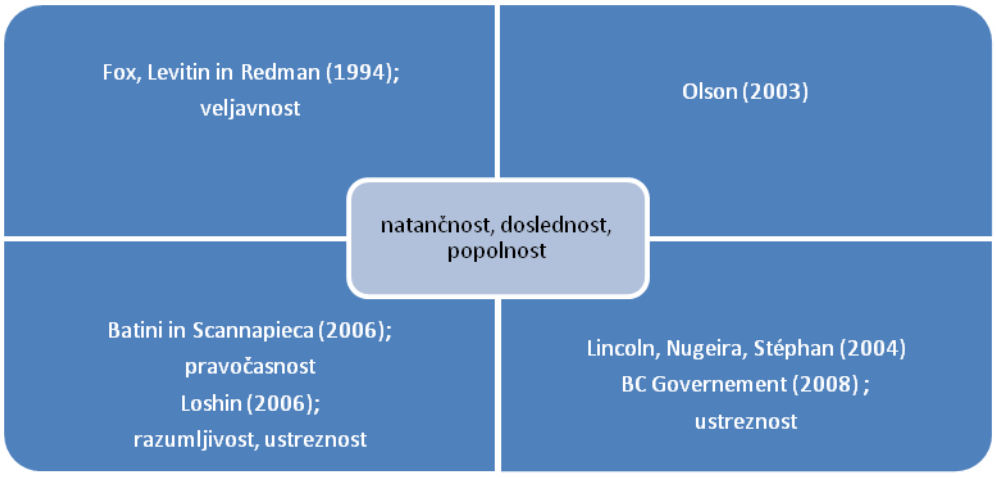
Uporabo enačbe je najlažje prikazati na praktičnem primeru. Vzemimo primer združbe, kjer se odločimo, da bomo kakovost podatkov o dolgu strank naši združbi merili kot kakovost treh razsežnosti, in sicer natančnosti, pravočasnosti in dostopnosti. V tem primeru je vrednost m = 3. Merjenje bomo izvršili samo v koraku shranjevanja podatkov, zato je vrednost n = 1. S subjektivno oceno, na primer anketiranjem uporabnikov, ugotovimo, da je natančnost ocenjena z vrednostjo 0,9, pravočasnost z 0,8 ter dostopnost 0,6. Kot poznavalci problematike se odločimo, da so ocene za nameravano uporabo realne, zato jih ne spreminjamo (vrednost U je pri vseh treh razsežnostih enaka 0). Če vrednosti vstavimo v enačbo, dobimo $Q = [K_1 * K_2 * \ldots K_n, \min(K_i)] = K_2$.

$K_2 = [0,9 * 0,8 * 0,6, 0,6] = [0,43, 0,6]$. Vidimo, da je kakovost za navedeni primer nekje med 0,43 ter 0,6, kar je slab rezultat. To pomeni, da imamo še ogromno možnosti za izboljšavo kakovosti podatkov.

8.5 Sklep 5: Izbor najpomembnejših razsežnosti kakovosti podatkov

V knjigi so proučevane naslednje razsežnosti kakovosti podatkov: natančnost, doslednost, popolnost, zaupanje v podatke, pravočasnost, ustreznost, razumljivost, dostopnost, za druge razsežnosti pa so navedene zgolj kratke opredelitve. To izhaja iz poudarjanja pomembnosti določenih razsežnosti posameznih avtorjev.

SLIKA 17: PRIKAZ NAJPOMEBNEJŠIH RAZSEŽNOSTI RAZLIČNIH AVTORJEV



Fox, Levitin in Redman (1994) poudarjajo, da so najpomembnejše razsežnosti kakovosti podatkov natančnost, doslednost, popolnost in veljavnost. Olson (2003) je v svojem delu izpostavil natančnost, omenja tudi doslednost in popolnost, medtem ko je druge dimenzije precej zanemaril. Loshin (2006) kot najpomembnejše razsežnosti izpostavi popolnost, ustreznost, doslednost, natančnost in razumljivost. Raziskava Lincolna, Nugeira in Stéphana

(2004) pa je pokazala, da so najpomembnejše razsežnosti kakovosti podatkov natančnost, doslednost, popolnost in ustreznost. Tej raziskavi pritrjuje tudi vlada Britanske Kolumbije (2008). Batini in Scannapieca (2006) pa izpostavljata natančnost, doslednost, popolnost ter pravočasnost.

S slike 17 je razvidno, da so najpomembnejše razsežnosti, ki so skupne omenjenim avtorjem, razsežnosti »natančnost«, »popolnost« in »doslednost«. Posledično je moje raziskovanje pri modeliranju in statistični raziskavi usmerjeno prav na te tri razsežnosti kakovosti podatkov, tj. natančnost, popolnost in doslednost.

8.6 Sklep 6: Opredelitev kakovosti izvedbe relacijskega podatkovnega modela

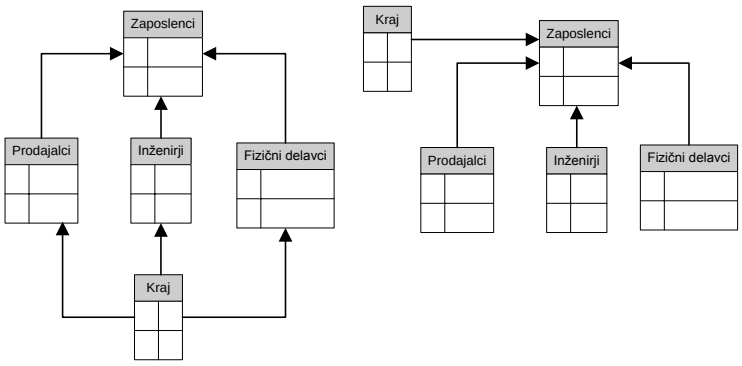
Poleg razsežnosti kakovosti podatkov obstajajo tudi razsežnosti kakovosti relacijske podatkovne sheme. Avtorji Ambler (2003), Redman (1996) ter Batini in Scannapieca (2006) jih v svojih delih podrobneje opredeljujejo. Redman (1996) opredeljuje šest glavnih razsežnosti ter 15 izvedenih razsežnosti kakovosti relacijske podatkovne sheme, ki jih Batini in Scannapieca (Batini in Scannapieca 2006) povzemata v sedem razsežnosti. Na tem mestu je treba poudariti, da obstaja razlika med podatkovno shemo, ki je dejansko logični načrt, ter FIRPM, ki je odvisen od posameznega sistema za ravnanje z zbirkami. Glavni, tuji ključi ter normalizacija so neodvisni od SUBP, podatkovni tipi ter deklarativne omejitve pa so odvisne od posameznega SUBP. Vendar vsi trije najpomembnejši ponudniki SUBP (Microsoft, Oracle in IBM) omogočajo podobne podatkovne tipe ter deklarativne omejitve. Zato se v knjigi izraza podatkovna shema ter relacijski model oziroma FIRPM uporabljata kot sopomenki.

Pravilnost modeliranja je prva v nizu razsežnosti in označuje, ali podatkovni model ustrezno modelira oziroma posnema stvarni svet. Omenjeno razsežnost je najlažje ponazoriti s praktičnim primerom. Predstavljajmo si entiteto »Zaposleni« z lastnostmi »#zaposlenega«, »priimek«, »ime«. Izvedba te entitete v podatkovnem modelu bi bila nepravilna, če bi jo razbili na dve preglednici in med njima vzpostavili povezavo 1 : 1. Takšna izvedba ne odseva realnega sveta, torej je nepravilna. V praksi se redko dogaja, da izvajamo povezave 1 : 1, vendar smo včasih prisiljeni v tovrstna dejanja zaradi fizičnih omejitev SUBP. Pravilnost izvedbe poslovnih pravil je druga v nizu razsežnosti kakovosti podatkovne sheme in opredeljuje, ali so poslovna pravila ustrezno izvedena v relacijskem podatkovnem modelu. Če ima združba zaposlene, ki so lahko tudi odgovorni za posamezen oddelek, pri čemer je posamezen zaposleni lahko odgovoren le za posamezen oddelek, potem je povezava med entitetama »zaposleni – oddelki« 1 : 1, kar je pravilno. Če bi bila izvedena povezava med navedenima entitetama 1 : N (ena proti mnogo), bi bila podatkovna shema nepravilna.

Pri načrtovanju relacijskega podatkovnega modela moramo paziti na podvajanje oziroma odvečnost sestavin. Po Redmanovem (1996) mnenju je podatkovna shema kakovostnejša, če se sestavine ne podvajajo oziroma moramo slediti cilju, da imamo v podatkovni shemi najmanjše mogoče število sestavin, tj. preglednic, povezav, polj, ki še odsevajo obravnavano problemsko področje. Najmanjše število dosežemo tedaj, ko ne moremo več odstraniti nobene sestavine brez vpliva na kakovost vsebine. Predpostavljajmo relacijski podatkovni model za spremljanje študentov pri različnih predmetih. Študent lahko obiskuje več predmetov (povezava 1 : N), posamezen predmet pa lahko poučuje tudi več učiteljev (povezava 1 : N). Če bi postavili neposredno povezavo tudi med študenti in učitelji (študenta lahko uči več učiteljev), je slednja povezava popolnoma odveč, saj slednjo vsebuje že povezava študent – predmet – učitelj.

Poleg navedenih razsežnosti kakovosti relacijske podatkovne sheme moramo paziti tudi na popolnost zasnove modela, s čimer opisujemo prisotnost vseh potrebnih lastnosti/atributov posamezne entitete, ki morajo ustrezati navedenim zahtevam. Če model ne vsebuje vseh lastnosti, potem je z vidika popolnosti nekakovosten, na primer entiteti »zaposleni« manjka lastnost »datum rojstva«.

SLIKA 18: PRIMER ZGOŠČENE PREDSTAVITVE PODATKOVNE SHEME



Vir: prirejeno po Batini in Scannapieca, Data quality: concepts, methodologies and techniques, 2006, str. 47.

S popolnostjo je povezana razsežnost primernost, s katero opisujemo, ali je model primeren ali ne, ali so načrtovalci pretiravali z modeliranjem ali ne. Načrtovalec lahko vgradi v shemo preveč podrobnosti, ki poslabšajo razumljivost sheme in nikakor ne izboljšajo njegove kakovosti.

Zadnji dve razsežnosti, ki po mnenju Batinija in Scannapieca (2006) vplivata na kakovost podatkovne sheme, sta preglednost oziroma razumljivost ter normalizacija. Vlogo normalizacije za kakovost podatkovne sheme obravnavajo tudi drugi, na primer Hussain, Shamail in Awais (2004).

Preglednost in razumljivost sta pravzaprav subjektivni merili in sta predvsem odvisni od posameznikovih sposobnosti dojemanja. Vpliv na preglednost ter razumljivost pa ima tudi

estetika predstavitve podatkovne sheme (Batini, Nardelli in Tamassia 1986 ; Sim 1996 ; Tammasia, Battista in Batini 1988). Batini, Nardelli, Tamassia, Battista in Sim priporočajo čim večjo simetrijo pri risanju podatkovne sheme, čim manjše število križanj črt, ki predstavljajo povezave, črte naj bodo ravne. Z upoštevanjem njihovih navodil se bomo izognili tako imenovanemu »špageti stilu« predstavitve podatkovne sheme. Poleg estetike pa na večjo razumljivost vpliva tudi zgoščenost (kompaktnost) predstavitve podatkovne sheme (Elmasri in Navathe 2000).

O normalizaciji je bilo v knjigi že veliko napisanega (razdelek 3.3). Poleg omenjenih normalnih form (prva, druga in tretja), poznamo še Boyce-Coddovo normalno formo ter četrto in peto. Med zadnjimi tremi normalnimi formami najbolj izstopa Boyce-Coddova normalna forma, ki označuje, da je povezava R v BCNF (Boyce-Coddovi normalni formi), kadar obstaja odvisnost med X -> Y v povezavi R in Y ni del X, X pa je kandidat za glavni ključ v povezavi R. Večina povezav, ki so v tretji normalni formi, so tudi v BCNF. Izjeme so le v naslednjih primerih (Zaiane 1998):

- kandidati za glavni ključ so sestavljeni ključi;
- v relaciji je več kot samo en kandidat za glavni ključ;
- kandidati za glavni ključ se prikrivajo, imajo skupne lastnosti.

V preglednici 13 je primer povezave, ki je v 3NF, ni pa v BCNF. Glavni ključ je sestavljen (#študenta, #učitelja), vendar je lahko glavni ključ tudi združek davčne številke študenta in #učitelja. V tem primeru predpostavljamo, da imamo vpisane izključno slovenske študente, ki imajo unikatno davčno številko, ki jih torej enolično določa. Za povezave, ki niso v BCNF, so značilne logične nedoslednosti, zato jih je treba skrbno obravnavati, kar pa ni velik problem, saj jih lahko hitro spremenimo tudi v BCNF (v navedenem primeru bi morali uporabiti le en združek polj, nikakor ne obeh, čeprav bi bilo najbolj pravilno, da bi davčno številko študenta prestavili v entiteto »študent«).

Piattini in Baroni ter njuni soavtorji (Baroni, Abreu in Calero 2005 ; Piattini, Calero in Genero 2001 ; Piattini et al. 2006) pa ocenjujejo kakovost relacijske sheme z vidika posamezne preglednice in celotne sheme. V preglednici obravnavajo število tujih ključev, število zapletenih podatkovnih tipov in »globino« povezav preglednice z drugimi preglednicami, v shemi pa obravnavata povezave med preglednicami ter zapletene podatkovne tipe. Opozarjajo na zahtevnejše vzdrževanje v primeru večjega števila zapletenih podatkovnih tipov, vendar njihov vpliv na kakovost podatkov prevlada nad morebitnimi negativnimi posledicami zahtevnejšega vzdrževanja.

PREGLEDNICA 13: PRIMER 3NF, KI NI BCNF

# študenta	Davčna številka študenta	#učitelja
123	12345678	987
231	87654321	654
213	78123456	587
213	78123456	654

Ambler (2006) pri opredeljevanju kakovosti relacijske sheme in njenem vplivu na kakovost podatkov opozarja na privzete vrednosti ter deklarativne omejitve. Trdi, da privzete vrednosti in število deklarativnih omejitev pozitivno vpliva na kakovost podatkov.

Naumann (2001) poudarja vlogo pravilnega načrtovanja poizvedb po podatkih za zagotavljanje kakovostnih podatkov. SUBP se pri načrtovanju poizvedbe odloča na podlagi indeksov ter po podatkovnih ključih, po katerih povezujemo med seboj preglednice, torej imajo po analogiji podatkovni ključi vpliv na kakovost podatkov. Načrtovanje glavnih in tujih ključev ima pomembno vlogo tudi pri neposrednih dostopih v zbirko podatkov. Predstavljajmo si, da moramo popraviti podatke v preglednici pozicija računa, vendar to lahko storimo samo s pomočjo preglednice glava računa. Če sta preglednici povezani med seboj prek enega ključa, na primer zaporedne številke računa, ki je enolična znotraj celotne preglednice glav računov, je poizvedba dokaj preprosta in možnost napak manjša. Če je ključ edinstven samo za določeno leto in določeno vrsto računa, je že potrebna večja pozornost, saj moramo v pogoj vključiti tri polja namesto enega, hkrati pa moramo paziti, da ima posamezna entiteta le en pripadajoč zapis v preglednici, česar pa umetno ustvarjeni ključi, na primer zaporedna številka zaposlenega, ne zagotavljajo. Pomembnost tujih ključev oziroma referenčne integritete poudarjata tudi English (2006) in Hay (2003), ki pravita, da morajo biti podatki s stališča referenčnih integritet brez napak. Iz navedenega primera je očitno, da je ustrezno načrtovanje glavnih in tujih ključev pomembno za kakovost podatkovnega modela.

Pri postavljanju glavnih in tujih ključev je treba skrbno izbirati polja/lastnosti/atribute, ki bodo opravljali vlogo ključev. Pri tem se postavlja vprašanje izbire med »naravnimi« in »umetnimi ključi«. Ko govorimo o »umetnih ključih«, po navadi mislimo na polja, kot so #zaposlenega, #študenta in ki jih določimo skozi programsko kodo. Na drugi strani imamo »naravne ključe«, na primer EMŠO, davčna številka. Vendar tudi ti ključi niso naravni, saj se z njimi ne rodimo, temveč nam jih umetno dodeli družba, v katero se rodimo. Prav zaradi slednjih pomislekov govorimo o »naravnih ključih« kot o lastnostih, ki jih uporabljamo tudi v stvarnem svetu. V svetu ne obstaja enotno mnenje o omenjeni problematiki, obstajajo le primeri najboljše prakse, ki govorijo v korist enih in drugih, torej je spet odvisno od problemske situacije. Logično je, če imamo mednarodno šolo, da EMŠO ter davčna številka nista primerna, saj tujci omenjenih lastnosti

nimajo vedno. V tem primeru bi uporabili umetni ključ, na primer vpisna številka študenta. Ko pa govorimo o registru državljanov Slovenije, sta davčna številka in EMŠO primerni lastnosti za glavni ključ. Pri izbiri ključev moramo upoštevati tudi značilnosti SUBP, predvsem z vidika hitrosti. Glede na osnovo računalnikov, ki razumejo le znaka 0 in 1, torej številki, je ravnanje s podatki precej hitrejše, če so ključi sestavljeni iz številk namesto iz znakov.

Zbirke podatkov torej vsebujejo dejstva, trditve iz resničnega sveta. Eden od najpomembnejših vidikov formalnih sistemov je logičnost, saj velja, da so logične razlike velike razlike (Wittgenstein, Pears in McGuinness 2001). Za podatkovni model je torej pomembno, da je logičen, hkrati pa mora zagotavljati popolnost – to pa je mogoče doseči samo z določenimi pravili. Pomembno je, da je pravil čim manj, da se z njimi strinjamo in da imajo preslikavo v stvarnem svetu. Katera pravila izbrati, kje in kako jih uvesti?

Pravila lahko opredelimo na različnih ravneh:

- podatkovnega tipa,
- preglednice (null, neponovljivost, deklarativne omejitve),
- podatkovne sheme,
- sprožilcev,
- shranjenih postopkov,
- programske kode na strani odjemalca,
- uporabniškega vmesnika.

Najstrožja so pravila na podatkovnem tipu, vendar imajo tudi slabost, da niso vedno razumljiva. Za primer vzemimo podatkovni tip celo število (angl. *integer*). Z izborom podatkovnega tipa smo določili dve stvari, in sicer kakšne znake lahko vnesemo v takšno polje ter kakšne operacije lahko izvajamo nad vnesenimi znaki. V polje »integer« lahko vnesemo samo številke oziroma natančneje cela števila. To polja omejuje tudi operacije, ki jih lahko izvajamo na posameznem tipu, v primeru tipa »integer« so to operacije seštevanja, odštevanja in množenja, z deljenjem pa so že težave, saj te operacije ne moremo vedno izvršiti. Navedeni primer kaže, da bližje izvoru postavimo pravilo, težje ga je zaobiti oziroma kršiti. Takšna pravila povzročajo težave pri razumevanju, še posebno pri zapletenih podatkovnih tipih, kar se kaže že pri datumskih poljih, kjer imamo mnogo standardov. Težave postanejo očitnejše pri lastnih podatkovnih tipih, kjer se pojavlja tudi težava z učinkovitostjo.

V relacijskem podatkovnem modelu potrebujemo dva tipa pravil, in sicer domenska pravila ter pravila, ki zagotavljajo povezave. Prav tako pa potrebujemo tudi glavne ključe, da zagotovijo edinstvenost zapisa.

Na podlagi proučene literature kakovost podatkovne sheme oziroma FIRPM določajo glavni in tuji ključi, deklarativne omejitve, podatkovni tipi, normalizacija/denormalizacija ter ustrezno modeliranje. V preglednici 14 so navedeni tudi avtorji, ki posamezno značilnost relacijskega podatkovnega modela posebej poudarjajo, čeprav je jasno, da so nekatere lastnosti med seboj odvisne, na primer pri procesu normalizacije se ukvarjamo tudi z določanjem glavnih in tujih ključev, vendar sta ti dve značilnosti obravnavani posebej, ker je normalizacija mogoča tudi brez njih.

PREGLEDNICA 14: OPREDELITEV RAZSEŽNOSTI KAKOVOSTI RELACIJSKE PODATKOVNE SHEME

Razsežnost kakovosti relacijske sheme	Avtorji, ki poudarjajo značilnost
Glavni ključi	(Maydanchik 2007 ; Naumann 2001)
Tuji ključi	(English 2006 ; Hay 2003 ; Maydanchik 2007 ; Naumann 2001)
Podatkovni tipi	(Baroni, Abreu in Calero 2005 ; Piattini, Calero in Genero 2001 ; Piattini et al. 2006)
Normalizacija/denormalizacija	(Batini in Scannapieca 2006 ; Hussain, Shamail in Awais 2004)
Deklarativne omejitve	(Ambler 2006)

Kakovost podatkovnega modela lahko izrazimo kot $M = k(P_n)$. (4)

P_n pomeni posamezno proučevano lastnost FIRPM.

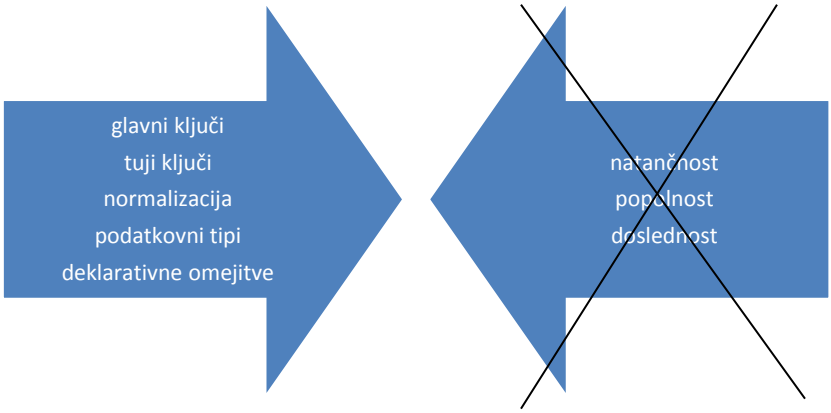
Navedene razsežnosti kakovosti relacijske podatkovne sheme niso edine, obstajajo še druge, ki z vidika obravnavane tematike v knjigi niso pomembne. Poleg subjektivnih razsežnosti, tj. preglednosti in razumljivosti, se pojavljajo tudi razprave o sprožilcih (angl. *triggers*) in shranjenih postopkih (angl. *stored procedure*). Te razsežnosti niso predmet obravnave, ker so že preveč »oddaljene« od temeljnih značilnosti podatkovnega modela, tj. stolpcev kot lastnosti entitet, preglednic kot entitet ter povezav med entitetami. Poudarjene značilnosti izhajajo iz osnovne teorije, kot jo je postavil Codd v 70. letih prejšnjega stoletja, sprožilci ter postopki pa so posledica evolucije SUBP ter izhajajo s področja objektnih zbirk podatkov.

8.7 Sklep 7: Proučevane lastnosti relacijskega podatkovnega modela in razsežnosti kakovosti podatkov

V podpoglavjih 8.5 ter 8.6 so bile izpostavljene lastnosti, ki so v nadaljevanju proučevane z modeliranjem z MS SQL 2005 in s statistično raziskavo. Kot proučevane lastnosti FIRPM so bili izpostavljeni glavni ključi, tuji ključi, podatkovni tipi, normalizacija/denormalizacija ter deklarativne omejitve. Pri razsežnosti kakovosti podatkov pa je bilo pokazano, da se avtorji

strinjajo, da so natančnost, popolnost in doslednost najpomembnejše razsežnosti kakovosti podatkov.

SLIKA 19: SMER MOREBITNE POVEZANOSTI OZIROMA VPLIVA MED FIRPM IN KAKOVOSTJO PODATKOV



FIRPM je shramba podatkov in se kot dogodek (A) vedno zgodi pred dogodkom (B), tj. kakovostjo podatkov. Po kavzalnem pristopu (angl. *causality*) to pomeni, da lahko samo dogodek (A) vpliva na dogodek (B), nasprotno pa ne. Torej, če obstaja povezava med dogodkoma (A) in (B), je slednja samo enosmerna in nikakor dvosmerna.

Naloga raziskovalnega dela je torej proučiti povezavo med Q in M, kjer Q pomeni kakovost podatkov, M pa kakovost FIRPM. Predpostavlja se naslednja povezava:

$Q = f(K)$, $K = (K_1, K_2, ..., K_n)$, $0 \le K_i \le 1, 0 \le Q \le 1$.

Ker naloga ocenjuje vpliv FIRPM na kakovost podatkov v PIS, torej v koraku povezovanja oziroma shranjevanja podatkov, sledi:

$Q = f(K_{2j}) = g(D_{2j}, U_{2j})$, pri čemer je $D_{2j} = k(P_n, O_m)$, (5)

kjer je P_n – lastnosti FIRPM (glavni ključi, tuji ključi, deklarativne omejitve, podatkovni tipi, normalizacija) ter O_m – drugi dejavniki (napačen vnos podatkov, nepravilno opredeljene entitete, sistemske napake).

Knjiga proučuje vpliv na kakovost podatkov izključno lastnosti FIRPM, zato druge dejavnike zanemarimo in predpostavimo:

$D_{2j} = k(P_n)$

V raziskavi torej proučujem, kako razsežnosti FIRPM glavni in tuji ključi, normalizacija, deklarativne omejitve in podatkovni tipi vplivajo na razsežnosti kakovosti podatkov, in sicer natančnost, popolnost ter doslednost.

9 RAZISKAVA

Raziskovanje v knjigi vsebuje vse sestavine, ki jih raziskava potrebuje, tj. proučevanje literature, zbiranje podatkov ter urejanje in analiziranje podatkov. Če te sestavine razčlenimo podrobneje, mora raziskava vsebovati naslednje sestavine: opredelitev raziskovalnega problema, načrtovanje raziskave, proučevanje literature, zbiranje podatkov, urejanje in analiziranje podatkov, razlaga rezultatov, razrešitev problema ter predstavitev ugotovitev (Kobeja 2001).

Nekaj korakov raziskave je bilo do zdaj že narejenih. V knjigi je opredeljen raziskovalni problem, tj. vprašanje vpliva izvedbe relacijskega podatkovnega modela na kakovost podatkov. Poleg problema je bila pregledana in proučena tudi ustrezna literatura, ki se dotika izbranega problemskega področja. Navedena literatura je bila osnova za sklepne trditve, ki so imele pomembno vlogo pri nadaljnjem poteku raziskave. Raziskava se je nadaljevala z zbiranjem podatkov, urejanjem in analiziranjem zbranih podatkov ter razlago dobljenih rezultatov. Slednji so temelj za razrešitev izbranega problema ter razpravo o odprtih vprašanjih.

Raziskovanje potrebuje tudi načrt raziskave, saj si ne moremo in smemo privoščiti napak, ker te navadno vodijo do pristranskih rezultatov, ki ne morejo popolnoma objektivno pripomoči k razrešitvi obravnavanega problema. Načrtovana raziskava je vsebovala opredelitev virov in tipov podatkov, metode zbiranje podatkov ter pripomočke za zbiranje in analiziranje podatkov.

Podatki so izvirni, dobljeni s pomočjo vprašalnika, ki so ga izpolnili v združbah, ki so oddale zaključni račun, iz vseh statističnih regij. Vzorec združb je bil naključen in nepristranski. Podatki so bili zbrani s pomočjo vprašalnikov, tj. anketiranja, ki spada med kvantitativne metode. Kvantitativne metode so se razvile na področju raziskovanja naravnih pojavov, pozneje pa so se razširile tudi na področje družboslovnih znanosti. Med kvantitativne metode spadajo anketiranje, laboratorijski poskusi, ekonometrija, matematično modeliranje itn. Kvalitativne metode imajo svoje izhodišče v dejstvu, da ljudi od ostalega živega sveta loči predvsem sposobnost govorjenja. Kvalitativne metode izhajajo s področja družboslovnih znanosti, mednje pa uvrščamo intervjuje, študije primerov, etnografske študije ipd. Namenjene so predvsem razlagi pojavov ter razmerij v družbi (Greenhalgh in Taylor 1997). Analiza s pomočjo kvalitativnih metod je načeloma bolj uporabna kot kvantitativna analiza, in sicer ko proučujemo povezave med ljudmi v neki združbi ali nekem procesu oziroma ko poskušamo dobiti podatke o zaznavah in razlagah ljudi (strokovnjakov) o neki storitvi, procesu ali problematiki. Podatki, ki jih dobimo prek kvalitativnih metod, so zelo uporabni, vendar jih je težko analitično obravnavati. Velikokrat je rečeno, da je zbiranje kvalitativnih podatkov preveč subjektivno, medtem ko je zbiranje kvantitativnih podatkov bolj predmetno in zanesljivo. To načeloma ne drži, saj so lahko kvantitativni podatki prav tako subjektivni (na primer izpraševalec izbira vprašanja po svoji

presoji, anketiranec pa odgovarja po svoji presoji). Za obe vrsti podatkov velja, da morajo biti čim bolj zanesljivi in preverljivi (Flere 2000).

Raziskovalci po navadi uporabljajo bodisi kvalitativne metode bodisi kvantitativne metode. V zadnjem času pa se vedno pogosteje uporablja združek kvalitativnih in kvantitativnih metod (Gable 1994 ; Mingers 2001). Z združkom metod so rezultati raziskave verodostojnejši. Metodo anketiranja pa sem uporabil zaradi želene delne primerjave z raziskavo TDWI (Russom 2006) in zaradi statistične obdelave, s katero sem želel potrditi glavno trditev. Poleg uporabe metode sem opravil modeliranje oziroma simulacijo s SUBP MS SQL 2005. Zamislil sem si dva razmeroma preprosta scenarija, ki nakažeta možnosti, ki jih FIRPM s sodobnim SUBP ponuja. Poleg logičnega načrtovanja vsebujeta scenarija tudi pripadajočo programsko kodo z namenom ponovljivosti modeliranja oziroma simulacije.

PREGLEDNICA 15: PRIMERJAVA RAZISKOVALNIH METOD

	Kvalitativne metode	**Kvantitativne metode**
Opredelitev vloge v družbi	Odnos	Struktura
Metode	Opazovanje, intervju	Poskus, anketa
Vprašanje	Kaj je X?	Koliko X-ov?
Analiza	Teoretična	Statistična
Moč metode	Preverjanje	Zanesljivost

Vir: Greenhalgh in Taylor, How to read a paper: papers that go beyond numbers (qualitative research), 1997.

V raziskovalnem delu je uporabljena kvantitativna metoda, ki je zajemala anketiranje ter posledično obdelavo anket in modeliranje oziroma simulacijo s pomočjo MS SQL 2005. Koraki izvajanja raziskave so bili naslednji:

- izdelava anketnega vprašalnika na podlagi teoretičnih spoznanj, raziskave TDWI in lastnih izkušenj, pri čemer je bil delež lastnih izkušenj največji, saj se z obravnavano problematiko ne ukvarja veliko raziskovalcev;
- modeliranje s pomočjo MS SQL 2005;
- preizkušanje vprašalnika na omejenem vzorcu;
- izvedba ankete na statistično nepristranskem vzorcu.

Anketiranje sem izvedel s spletnim vprašalnikom, saj omogoča ekonomsko učinkovito anketiranje ter zmožnost hitre in pregledne obdelave rezultatov. Poleg tega večina spletnih orodij za anketiranje omogoča neposreden prenos podatkov v Excel in SPSS za nadaljnje podrobnejše obdelave.

9.1 Opredelitev virov in tipov podatkov raziskave

V raziskavah lahko uporabljamo izvirne ali/in drugotne podatke. Drugotni podatki so z vidika gospodarnosti ustreznejši, saj prihranijo čas in denar, vendar hkrati niso vedno primerni za reševanje raziskovalnega problema. Poleg tega je točnost drugotnih podatkov težje preverjati v primerjavi z izvirnimi podatki. Kljub naštetim slabostim se je vredno potruditi in preveriti, ali za izbran raziskovalni problem obstajajo drugotni podatki ali ne.

Za raziskovalni problem, tj. vpliv kakovosti izvedbe relacijskega podatkovnega modela na kakovost podatkov, drugotni podatki ne obstajajo, zato je bila edina možnost zbiranje izvirnih podatkov. Izvirne podatke sem zbral s pomočjo vprašalnika, v katerem so anketiranci izrazili svoja mnenja, znanje in odnos do izpostavljene problematike.

Zaradi narave raziskovalnega problema in predpostavke o znanju izvedbe relacijskega podatkovnega modela v posameznih združbah so bili anketiranci iz dveh skupin, in sicer slovenske združbe kot uporabniki PIS ter ponudniki PIS v prej omenjenih slovenskih združbah.

Raziskava je morala biti nepristranska, zato so bile zajete združbe različnih panog in velikosti iz vseh statističnih regij. V vzorec populacije niso bili vključeni samostojni podjetniki, niti združbe, ki niso oddale zaključnega računa.

Razlogov, da sem se odločil za takšen vzorec, je več. Iz lastnih izkušenj izhaja, da samostojni podjetniki redko uporabljajo poslovno obveščanje, čeprav se tudi na tem področju kaže trend razvoja preprostih orodij. Zaradi tega problemi s kakovostjo podatkov niso poudarjeni. Njihova dejavnost v smislu uporabe PIS je v pretežni meri usmerjena v izdane in prejete račune ter v poravnavo obveznosti do države. Poleg tega samostojni podjetniki navadno nimajo lastnega PIS, temveč to dejavnost zanje opravljajo računovodski servisi. Ta dejavnost pa se s pojavom spletnih računovodskih rešitev seli na svetovni splet. V Sloveniji imamo dva velika ponudnika spletnega računovodstva, miniMax ter eRacuni.

Anketiranci v združbah, ki so uporabniki PIS, so bili ključne osebe v službi za informatiko oziroma osebe, ki imajo neposredno povezavo s PIS. Za razreševanje raziskovalne problematike je bilo najpomembnejše poznavanje težav, ki izhajajo iz kakovosti podatkov, predvsem težav, ki jih imajo uporabniki PIS, vzrokov ter načinov reševanja omenjenih težav.

Za razjasnitev raziskovalnega problema bi bilo najboljše, če bi anketiranci v združbah dobro poznali tudi PIS z vidika podatkovnega modela. Vendar sem se popolnoma zavedal, da je to prej izjema kot pravilo. Če anketiranci v združbah niso znali odgovoriti na vsa vprašanja, sem podatke poskušal dobiti neposredno od ponudnikov PIS.

V združbah, ki so ponudniki PIS, so bili zaželeni anketiranci osebe, ki neposredno sodelujejo pri razvoju PIS. Najbolj zaželene so bile osebe, ki imajo celoten pregled nad zgradbo podatkovnega modela. Iz izkušenj sem predpostavljal, da imajo le redki posamezniki celovit pregled nad celotnim podatkovnim modelom.

Poleg mnenj, znanja in odnosa do kakovosti podatkov ter podatkovnega modela sem zbral tudi osnovne opisne podatke, ki so pomagali razjasniti raziskovalni problem še z dodatnega vidika.

Poleg anketiranja sem s pomočjo SUBP MS SQL 2005 opravil tudi modeliranje podatkovnega modela na dveh preprostih scenarijih, s katerima sem izpeljal model, katere lastnosti FIRPM vplivajo na natančnost, doslednost in popolnost. Tako izpeljani model sem poskušal potrditi tudi s statistično analizo.

Vrstni red raziskave in posledično poskusov dokazovanja pravilnosti izvedenih trditev, je bil naslednji:

- modeliranje s SUBP MS SQL 2005,
- anketiranje in analiza anket.

9.2 Vidik relacijskega podatkovnega modela

Gradnja hiše se začne s postavitvijo dobrih temeljev, ki omogočajo hiši dolgotrajen obstoj. Dobri temelji so zagotovilo, da bo hiša obstala tudi ob različnih naravnih pretresih. Ko postavimo temelje, lahko začnemo graditi hišo, ji dodamo funkcije, ki jih pozneje lahko še dodatno dopolnjujemo. Po določenem času, ko obstoječo hišo načne zob časa, jo lahko porušimo do temeljev in na njih zgradimo novejše bivališče. Izraz porušiti hišo do temeljev je zelo primeren tudi v povezavi s kakovostjo podatkov. Če imamo dobre temelje pri hiši, na njih lahko zgradimo dom za več generacij, prav tako kot lahko na dobrem podatkovnem modelu zgradimo več generacij uporabniških rešitev. Med gradnjo hiše in podatkovnim modelom lahko potegnemo še eno vzporednico. Ko poslušam ljudi, kaj se jim pri nakupu nepremičnin zdi pomembno, vsi govorijo le o funkcijah, primerni lokaciji in ceni, nihče oziroma redki pa se pozanimajo tudi o temeljih nepremičnin, stenah, materialu. Tudi pri nakupu PIS združbe delujejo podobno. Veliko pozornosti posvečajo funkcijam uporabniških rešitev, videzu in ceni, redke pa so združbe, ki pregledajo tudi temelje uporabniške rešitve, tj. podatkovni model. S tem si napravijo medvedjo uslugo, saj lahko izberejo rešitev, ki ima slab podatkovni model, in tako ogrozijo svoje premoženje, tj. podatke. Združbe se morajo zavedati, da so podatki, poleg vrednot in kakovosti organizacije, njihovo najpomembnejše premoženje. Mnogi bi tej trditvi ugovarjali in zagovarjali stališče, da so najpomembnejše premoženje združb zaposleni ter njihovo znanje, vendar se je

treba zavedati, da zaposleni prihajajo in odhajajo, nekateri celo umrejo, podatki pa združbam ostajajo, zato jim je treba nameniti ogromno pozornosti.

Kako pomembno postaja področje kakovosti podatkov, nakazujejo tudi številna dejanja Microsofta, ki je ponudil uporabniško rešitev za področje MDM (*Master Data Management*). Tej manjka še ogromno funkcij, da bi ujela vodilne uporabniške rešitve na tem področju, na primer *Data Flux*. Hkrati pa me skrbi Microsoftova dejavnost v zvezi z željo približevati razvoj podatkovnega modela ljudem, katerih osnovna usmeritev ni podatkovno modeliranje. Počasi, vendar vztrajno prestavljajo funkcije podatkovnega modeliranja v *Visual Studio*, ki je namenjen razvijalcem vmesnih in odjemalčevih ravni. Njihova naravnanost je najbrž predvsem posledica tržne naravnanosti po čim širši uporabi njihove uporabniške rešitve, hkrati pa najbrž poskušajo ugoditi tudi različnim pritiskom oziroma željam po neki novi podatkovni strukturi, ki ni relacijska. Kot v panogi gradnje hiš se tudi v svetu razvijanja uporabniških rešitev pojavljajo različne modne smernice, ki jim velike združbe poskušajo ugoditi. Microsoft je leta 2007 začel pogosto uporabljati izraz »*beyond relational*«, s katerim opisuje nove, zapletene podatkovne tipe (na primer prostorski podatkovni tip, listinski podatkovni tip), vendar je izraz »*beyond relational*« popolnoma napačno uporabljen. Relacijski podatkovni model namreč nikjer ne predpisuje, katere podatkovne tipe lahko uporablja, prav tako nikjer ne zapoveduje načina fizične izvedbe. Oba omenjena vidika sta izključno v rokah ponudnikov posameznih SUBP.

Zapostavljanje relacijskega podatkovnega modela se mi ne zdi premišljena poteza. Relacijski podatkovni model temelji na znanosti, tj. teoriji množic. In stvari, ki temeljijo na znanosti, imajo daljši rok veljavnosti. Poleg tega sodobni SUBP ponuja ogromno pravil, ki zagotavljajo celovitost podatkov, torej varujejo najpomembnejše premoženje združb. Bolj ohlapna bodo pravila, manj varovani bodo naši podatki. Navsezadnje lahko združbe hranijo podatke v Excelovih preglednicah, vendar je njihova kakovost v daljšem časovnem obdobju vprašljiva.

Ker relacijski podatkovni model temelji na znanosti, bo najbrž še nekaj časa ostal glavni podatkovni model za večino uporabniških rešitev. Poleg tega ponudniki SUBP nenehno dopolnjujejo svoje uporabniške rešitve z dodajanjem novih funkcij, kar relacijskemu modelu podaljšuje njegovo življenjsko dobo. Prav zaradi pomembnosti temeljev uporabniške rešitve bi morali slovenski in tudi tuji razvijalci temeljem nameniti več pozornosti, saj s tem omogočajo kakovostnejše uporabniške rešitve ter možnost uporabe temeljev daljše časovno obdobje. Navsezadnje ni pomembno, ali za izdelavo poslovne logike in uporabniškega vmesnika uporabimo *FoxPro, C#, VB.NET* in njihove novejše različice, če je podatkovni model dovolj kakovosten. Ustrezen podatkovni model nam omogoča tudi lažji razvoj in dodajanje funkcij, kar posledično znižuje stroške razvijalcem PIS.

Veliko težav ponudnikov PIS se pojavlja zaradi samodejnega prehoda iz zastarelih datotečnih sistemov v nova razvojna okolja. Najboljši primer za to je uporabniška rešitev Navision, pri kateri pa je treba upoštevati čas, v katerem je nastala, saj takrat sodobni SUBP še ni bil na voljo in so Danci razvili svojo zbirko podatkov. Podobno je s SAP, ki se je razvil iz glavne knjige in materialnega poslovanja, danes pa vsebuje približno 40.000 preglednic. Poseg združbe, ki uporablja SAP, neposredno v zbirko podatkov je običajno prepovedan. Nekoč sem se mudil v znani večji slovenski združbi, kjer zaradi omenjene prepovedi vlada veliko nezadovoljstvo med uporabniki PIS. Če odmislim ekonomsko naravnanost uvajalcev SAP, ki za tovrstne posege zaračunavajo velike vsote denarja, je takšna prepoved smiselna zaradi ohranjanja podatkovne celovitosti. Uvajalec SAP tej združbi ne zagotavlja ustreznega delovanja PIS v primeru njihovega neposrednega dostopa do podatkov. Uporabniki se počutijo ogoljufane, saj so podatki njihova last. Vendar v tem primeru kakovost podatkov zagotavljajo le z administrativno prepovedjo neposrednega dostopa do podatkov. Organizacijski predpisi lahko v veliki meri nadomestijo slabšo kakovost podatkovnega modela, vendar to ne sme biti izgovor za slabše podatkovno modeliranje.

Preverjanje poslovnih in podatkovnih pravil je treba zagotoviti na več ravneh, kajti če ena od ravni odpove, ostanejo še druge, ki neustrezen podatek ustrezno obravnavajo. In podatkovni model je zadnja in najpomembnejša raven, ki mora preprečevati vpis neustreznih vrednosti, še posebno v tem času, ko različna orodja za poslovno obveščanje zavzemajo pomembno vlogo v odločitvenih procesih. Izraz GIGO s stališča kakovosti podatkov lahko spremenimo v izraz GIwGO (*Garbage In – worse Garbage out*), saj nekakovostni podatki, če ni zdravega razuma, lahko pripeljejo do resnično nespametnih odločitev. Vzrok za takšno pojmovanje lahko hitro najdemo v sestavi uporabniških rešitev. Razširjenost analitičnih orodij je velika in veliko sodobnih uporabniških rešitev vsebuje OLAP že v svoji standardni različici. Ker je OLAP sodoben, moden, ga uporablja tudi veliko odločevalcev, ki lahko hitro, v prevelikem navdušenju, pozabijo na osnovno sestavino OLAP-a, tj. podatke oziroma njihovo kakovost.

Če proučimo razmerje med vlogo kakovosti podatkov ter različnimi analizami, lahko hitro vidimo, da vloga kakovosti podatkov narašča z zapletenostjo analiz. Omenjeno razmerje lahko dokažemo preprosto s prikazom praktičnega primera. Predpostavimo, da imamo združbo, ki prodaja kolesa in pripadajočo opremo, usmerjeno predvsem za delo v vinogradih. V šifrantu krajev imamo napačno vnesen kraj Šempeter. Dejansko imamo v mislih Šempeter pri Gorici, vnesli pa smo ga kot Šempeter pri Celju, torej je drugo geografsko območje. Podatek je sintaktično natančen, semantično pa ne. Da bi posledice takšne napake lažje razumeli, jo bomo obravnavali z vidika SSOP, OLAP in SZOP. Ko napako obravnavamo z vidika SSOP, posledice niso veliko oziroma jih morda sploh ni. Če na primer izdamo račun za kupca iz Šempetra pri Gorici, se napaka nikjer ne pokaže, saj geografsko območje pri izdanih računih po navadi ni navedeno.

Združba tako lahko mirno posluje. Težave bi v obravnavanem primeru verjetno nastopile po izdelavi skladišča podatko,v nadgrajenega s sistemom OLAP. Geografsko območje se po lastnih izkušnjah pogosto znajde na spisku želja hierarhije razsežnosti kupcev posameznih združb v OLAP. Zaradi storjene napake bi vse kupce iz kraja Šempeter pri Gorici preselili v geografsko območje Celje z okolico, zato bi bili podatki napačni (v praksi bi takšno napako verjetno hitro odkrili, vendar bom to dejstvo zaradi namena dokazovanja odvisnosti med kakovostjo podatkov in zapletenostjo analiz zanemaril). Rezultati analize OLAP bi bili v tem primeru popolnoma napačni in bi odločevalca zmedli. Če bi bilo na primer ogromno prebivalcev Šempetra pri Gorici kupcev koles, bi bila dejanska prodaja koles na Goriškem velika, vendar bi v našem analitičnem sistemu to prikazovalo kot prodajo na območju Celja z okolico. Omenjeno dejstvo bi management združbe lahko napeljalo k dejavnostim pospeševanja prodaje na Goriškem, čeprav bi bilo to bolj smiselno storiti na Celjskem. Še pomembnejša je kakovost podatkov pri uporabi SOZP. S tehniko »drevesa odločitev« bi opravili analizo podatkov in hipotetično ugotovili naslednje:

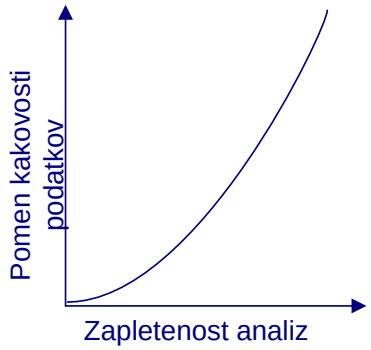
Verjetnost, da posameznik, ki je lastnik hiše, živi blizu središča kraja, nima avtomobila in je z geografskega območja Celje z okolico, kupi kolo, je 86 %.

Z obstoječim spoznanjem bi v trženjske dejavnosti zajeli le možnostne kupce, ki ustrezajo vzorcu, odkritem s SOZP. V veri, da bomo z nameravano trženjsko dejavnostjo povečali dobiček, dosežemo nasprotni učinek, tj. zmanjšanje dobička, kajti pravilna ugotovitev se glasi:

Verjetnost, da posameznik, ki je lastnik hiše, živi blizu središča kraja, nima avtomobila in je z geografskega območja Goriško, kupi kolo, je 86 %.

Iz navedenega primera je razvidno, da z vse bolj zapletenimi analizami narašča vloga kakovosti podatkov (slika 20).

SLIKA 20: ODVISNOST MED KAKOVOSTJO PODATKOV IN ANALIZAMI



Čeprav na tem mestu še ne morem z gotovostjo trditi, da kakovosten podatkovni model vpliva na kakovostne podatke, pa lahko oblikujem trditev, da nekakovostni podatkovni model omogoči pogoje za nastanek nekakovostnih podatkov. Zdravljenje, torej odpravljanje nekakovosti podatkov je po analogiji z uradno medicino slabša možnost kot preprečevanje nekakovosti. Ker ima relacijski model v svojem standardu opredeljenih kar nekaj pravil, ki omogočajo ustrezno modeliranje, je ta pravila treba upoštevati, seveda z občutkom za nameravano uporabo. Pri načrtovanju podatkovnega modela je treba prepoznati ključne entitete (množice) in povezave med njimi ter jim nameniti veliko pozornosti pri preslikavi v digitalni svet. Tega procesa se morajo lotiti posamezniki, ki imajo sposobnost abstraktnega razmišljanja, vsekakor pa se ga ne smejo lotevati začetniki. Te problematike se morajo zavedati tudi združbe, ki razvijajo PIS. Napačno zasnovan podatkovni model ima lahko hude posledice za uporabniško rešitev in s tem za ugled in stroške združbe, zato mora biti podatkovno modeliranje v rokah usposobljenih ljudi. Hkrati z omenjenim zavedanjem pa se bo moralo spremeniti tudi zavedanje odjemalcev PIS. Zunanji videz, v tem primeru videz uporabniškega vmesnika, pač ni zagotovilo za kakovost celotne uporabniške rešitve, še posebno njenega podatkovnega modela. Pri tem mi prihaja na misel primerjava med japonskimi in evropskimi avtomobili. Japonski avtomobili še danes oblikovno niso kos evropskim avtomobilom, po kakovosti in zanesljivosti pa jih močno prekašajo. V zavest združb se bo moralo vtisniti dejstvo, da pri odločitvi za nakup določenega PIS podatkovni model igra pomembno, če ne celo ključno vlogo.

9.3 Modeliranje s SUBP – MS SQL 2005

Z modeliranjem sem postavil model, kako lastnosti oziroma razsežnosti FIRPM vplivajo na posamezno proučevano razsežnost kakovosti podatkov. V nadaljevanju sem pokazal, kako imajo normalizacija, ključi, deklarativne omejitve ter podatkovni tipi možnost, da zagotovijo kakovostnejše podatke. **Z modeliranjem sem izpeljal model, ki sem ga pozneje poskušal potrditi s statistično analizo.** Za dokazovanje vpliva FIRPM na doslednost sem upošteval Batinijevo in Scannapiecovo (2006) opredelitev doslednosti, ki pravi, da je doslednost kot razsežnost kakovosti podatkov usklajenost podatkov s semantičnimi pravili, pri popolnosti pa sem proučeval vpliv FIRPM na manjkajoče vrednosti. Model prikazuje tudi vpliv FIRPM na natančnost podatkov.

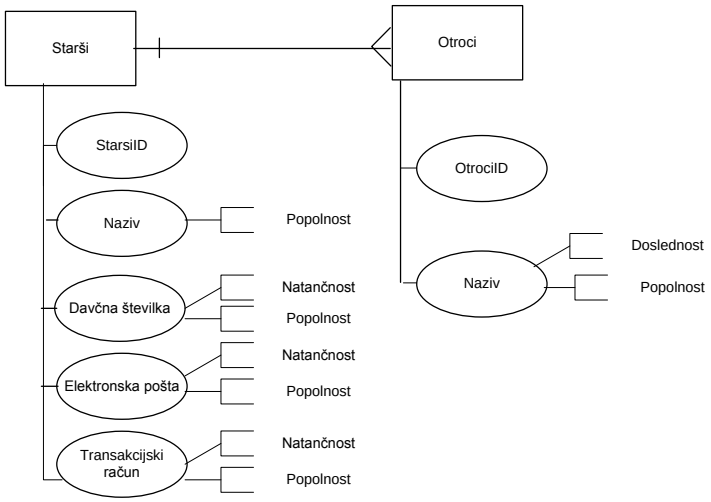
9.3.1 Načrtovanje podatkovnega modela – scenarij 1: Vrtec

Scenarij: Sem načrtovalec zbirke podatkov za vrtec, v katerega sprejmejo največ tri otroke posameznih staršev. V zbirki želi vrtec hraniti ime staršev, davčno številko staršev, imena otrok, elektronsko pošto staršev, saj jih želi obveščati o pomembnih dogodkih v vrtcu, in transakcijski račun enega od staršev zaradi izstavitve računa. Pri obravnavi scenarija predpostavljamo, da

podatke vnašamo neposredno v preglednice, tj. brez uporabe uporabniškega vmesnika. Najprej je treba narediti razširjen entitetno-relacijski diagram z vidikom kakovosti (slika 21).

Entitetno-relacijski diagram je logična zasnova podatkovnega modela, ki je neodvisna od fizične izvedbe. Po logičnem modelu lahko preidem na načrtovanje fizičnega modela. Z naslednjima stavkoma TSQL naredim dve preglednici, preglednico »Starsi« in preglednico »Otroci«:

SLIKA 21: RAZŠIRJENI ENTITETNO-RELACIJSKI DIAGRAM Z VIDIKOM KAKOVOSTI – SCENARIJ 1




```
CREATE TABLE Starsi (
  ZaporednaStevilka  INT PRIMARY KEY,
  NazivStarsa        VARCHAR(300),
  ElektronskaPosta   VARCHAR(100),
  TransakcijskiRacun VARCHAR(15))
GO
CREATE TABLE Otroci (
  ZaporednaStevilka       INT,
  ZaporednaStevilkaStarsa INT,
  NazivOtroka             VARCHAR(300),
  CONSTRAINT PK_Otroci PRIMARY KEY(ZaporednaStevilka,ZaporednaStevilkaStarsa))
```

V preglednici »Starsi« sem za glavni ključ določil polje »ZaporednaStevilka«. Glavni ključ ne dovoljuje ponavljajočih se vrednosti, ker je njegova naloga enolično določati zapis v preglednici. S tem sem zmanjšal možnost napak, nisem pa jih odpravil. Nič namreč ne preprečuje vnosa istih staršev pod drugo zaporedno številko.

Težave lahko nastanejo tudi pri vnosu imena in priimka. Iz naziva polja ni razvidno, ali moramo najprej vnesti ime ali priimek, kar mi daje slutiti, da predlagani model ni najboljši. Veliko boljše bi bilo, da bi imeli ločeni polji za ime in priimek, saj se s tem izboljša verjetnost za semantično

natančnost podatka, kajti če je v navadi, da najprej navajamo ime, potem je podatek »Novak Janez« sicer sintaktično natančen, semantično pa nenatančen.

Enak problem obstaja tudi v preglednici »Otroci«.

Z naslednjim ukazom izdelamo preglednici, ki zagotavljata višjo kakovost podatkov:

```
CREATE TABLE Starsi (
  ZaporednaStevilka  INT PRIMARY KEY,
  Ime                VARCHAR(150),
  Priimek            VARCHAR(150),
  DavcnaStevilka     VARCHAR(9),
  ElektronskaPosta   VARCHAR(100),
  TransakcijskiRacun VARCHAR(15))
GO
CREATE TABLE Otroci (
  ZaporednaStevilka       INT,
  ZaporednaStevilkaStarsa INT,
  ImeOtroka          VARCHAR(150),
  PriimekOtroka      VARCHAR(150),
  CONSTRAINT PK_Otroci PRIMARY KEY(ZaporednaStevilka,zaporednaStevilkaStarsa))
```

Navedeni model ima še vedno veliko slabosti, na primer polje elektronska pošta. Elektronska pošta ne sme vsebovati določenih znakov, mora pa vsebovati @. Vnos v znakovno polje ne omogoča nobenega preverjanja, zato je možnost napak velika. Podobno je s poljem »TransakcijskiRacun«. Transakcijski račun ima znano sestavo, in sicer je sestavljen iz petih številk, vezaja ter naslednjih desetih številk. Zadnja številka je kontrolna številka, kar omogoča visoko stopnjo sintaktične in semantične natančnosti. Trenutno stanje, tj. znakovno polje, ne omogoča nobenega preverjanja. Trenutno stanje prav tako ne zahteva vnosa vrednosti v nobeno polje, razen v polja, ki predstavljajo glavne ključe. Da bi zagotovil popolnost z vidika vrednosti NULL/NOT NULL ter onemogočil prazne vrednosti, moram uporabiti deklarativne omejitve.

```
CREATE TABLE Starsi (
 ZaporednaStevilka  INT PRIMARY KEY,
 Ime                VARCHAR(150) NOT NULL,
 Priimek            VARCHAR(150) NOT NULL,
 DavcnaStevilka     VARCHAR(9)   NOT NULL,
 ElektronskaPosta   VARCHAR(100) NOT NULL,
 TransakcijskiRacun VARCHAR(15)  NOT NULL)
GO
CREATE TABLE Otroci (
 ZaporednaStevilka       INT,
 ZaporednaStevilkaStarsa INT,
 ImeOtroka          VARCHAR(150) NOT NULL,
 PriimekOtroka      VARCHAR(150) NOT NULL,
 CONSTRAINT PK_Otroci PRIMARY KEY(ZaporednaStevilka,zaporednaStevilkaStarsa));
GO
ALTER TABLE Starsi
```

```
ADD CONSTRAINT Ime_NI_Neznano CHECK (Ime<>''),
CONSTRAINT Priimek_NI_Neznan CHECK (Priimek<>'')
CONSTRAINT Priimek_NI_Neznan CHECK (TransakcijskiRacun<>'')
GO
ALTER TABLE Otroci
ADD CONSTRAINT Imeotroka_NI_Neznano CHECK (ImeOtroka<>''),
CONSTRAINT PriimekOtroka_NI_Neznan CHECK (PriimekOtroka<>'')
```

S tem sem zagotovil, da bodo podatki popolnejši z vidika vrednosti NULL/NOT NULL (vsa polja) in praznih vrednosti (imena in priimki staršev ter otrok). Postavljene deklarativne omejitve vnašalca prisilijo, da vnese imena in priimke staršev ter otrok in transakcijski račun staršev. Polje za vnos podatka o elektronski pošti ne dopušča vrednosti NULL, dopušča pa prazno vrednost. Vendar še zmeraj nisem zagotovil nadzora nad vnosom v polji »ElektronskiNaslov« in »TransakcijskiRacun«. V tem primeru imam dve možnosti, in sicer uporabo deklarativnih omejitev ali zapletenih podatkovnih tipov. V nadaljevanju sta predstavljeni obe možnosti, saj bom za polje »ElektronskiNaslov« izdelal lastni podatkovni tip, za »TransakcijskiRacun« pa deklarativno omejitev.

Lastni podatkovni tip »ElektronskiNaslov« je v navedenem primeru ustvarjen v jeziku C#:

```
using System;
using System.Data;
using System.Data.SqlClient;
using System.Data.SqlTypes;
using Microsoft.SqlServer.Server;
using System.Text;
using System.Text.RegularExpressions;
using System.IO;

[Serializable]
[Microsoft.SqlServer.Server.SqlUserDefinedType(Format.UserDefined, MaxByteSize = 8000)]
public struct EmailCS : INullable, IBinarySerialize
{
 //Preverjanje elektronske pošte, ki ne ustreza strukturi nekdo@nekazdruzba.domena
 private static readonly Regex RegExParser = new Regex(@"^([\w-]+\.)*?[\w-]+@[\w-]+\.([\w-
]+\.)*?[\w]+$", RegexOptions.CultureInvariant);

 // Notranja metoda za elektronski naslov
 private StringBuilder parsedemail;
 // Ali je vrednost NULL?
 private bool m_Null;
 // Obravnava vrednosti
 public EmailCS(string value)
   {
 this.parsedemail = new StringBuilder();
 this.parsedemail.Append(value);
 this.m_Null = false;
   }
```

```
// Privzeta metoda
public override string ToString()
{
 return this.parsedemail.ToString();
}

// Obravnava NULL vrednosti
public bool IsNull
{

get
 {
  return m_Null;
 }
}

public static EmailCS Null
{
 get
 {
  EmailCS h = new EmailCS();
  h.m_Null = true;
  return h;
 }
}

// Privzeta metoda
public static EmailCS Parse(SqlString s)
{
 if (s.IsNull)
  return Null;

 // Ali je vnešena vrednost ustrezna?
 string value = s.ToString();
 Match m = RegExParser.Match(value);

 // Vnesena vrednost je nepravilna ali predolga.
 if (!m.Success || value.Length > 4000)
  throw new ArgumentException(
   "Nepravilna struktura naslova elektronske pošte. "
   + "Struktura nekdo@nekazdruzba.domena ima lahko največjo dolžino 4000 znakov.");

 // Potrdimo ustreznost
 EmailCS u = new EmailCS(value);
 return u;
}

// Serealizacija
public void Read(BinaryReader r)
{
```

```
 parsedemail = new StringBuilder(r.ReadString());
}

public void Write(BinaryWriter w)
{
 w.Write(this.parsedemail.ToString());
}
}
```

Deklarativna omejitev za polje »TransakcijskiRacun« uporablja funkcijo »dbo.KontrolaTRR«, katere videz je naslednji:

```
CREATE FUNCTION [dbo].[KontrolaTRR](@Vrednost VARCHAR(19))
RETURNS VARCHAR(50) AS
BEGIN
   DECLARE @Napaka INT;
   DECLARE @Dolzina int
   DECLARE @Levi CHAR(5);
   DECLARE @Desni CHAR(10);
   DECLARE @Vezaj CHAR(1);
   DECLARE @KontrolnaVrednost INT;

   SET @Napaka  = 1
   SET @Dolzina = LEN(@Vrednost)
   SET @Levi    = LEFT(@Vrednost, 5)
   SET @Desni   = RIGHT(@Vrednost, 10)
   SET @Vezaj   = SUBSTRING(@Vrednost, 6, 1)

IF NOT( (@Dolzina=16) AND (ISNUMERIC(@Levi)=1) AND  (ISNUMERIC(@Desni)=1) AND
(@Vezaj='-'))
   SET @Napaka = -1
ELSE
BEGIN
    SET @KontrolnaVrednost = 98 - (CONVERT(bigint, @Levi+LEFT(@Desni,8)+'00')
% 97)
    IF (@KontrolnaVrednost<>CONVERT(bigint,RIGHT(@Vrednost, 2)))
        SET @Napaka = -1
END


RETURN @Napaka
END
```

Funkcijo »dbo.KontrolaTRR« uporabim na naslednji način:

```
ALTER TABLE Starsi
CONSTRAINT TRR_Kontrola  CHECK (dbo.KontrolaTRR(TransakcijskiRacun)=1)
```

S tem sem zagotovil visoko stopnjo sintaktične in semantične natančnosti. Pravzaprav je edina možnost za nekakovosten podatek vnos sintaktično pravilnega transakcijskega računa druge osebe. Če bi se hotel izogniti temu dejanju, bi moral preveriti podatke pri Banki Slovenije.

Vendar pa je treba upoštevati vidik uporabe in v navedenem primeru ter tudi v večini drugih je izvedbeni način polja »TransakcijskiRacun« več kot zadovoljiv.
V preglednico »Starsi« lahko večkrat vnesemo iste starše. Enako je v preglednici »Otroci«. Odprava težave je mogoča z izvedbo polja »DavcnaStevilka« v preglednici »Starsi« in »DavcnaStevilka« v preglednici »Otroci«. Ker v vrtec lahko pridejo tudi starši in otroci, ki nimajo slovenske davčne številke, ne morem uporabiti podobne deklarativne omejitve kot v primeru transakcijskega računa, preverjal pa bomo njeno edinstvenost.

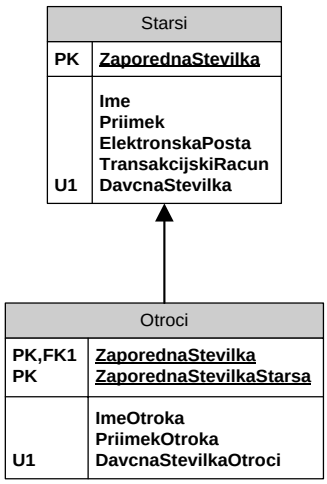
Z navedenim ukazom sem dodal polji »DavcnaStevilka« in »DavcnaStevilkaOtroka« ter zagotovil njeno edinstvenost:

```
ALTER TABLE Starsi
ADD DavcnaStevilka VARCHAR(20) NOT NULL DEFAULT ''
CONSTRAINT Davcna_Edinstvenost  UNIQUE NONCLUSTERED
GO
ALTER TABLE Otroci
ADD DavcnaStevilkaOtroci VARCHAR(20) NOT NULL DEFAULT ''
CONSTRAINT DavcnaOtroci_Edinstvenost  UNIQUE NONCLUSTERED
```

Podatkovni model, izveden do zdaj, omogoča visoko stopnjo sintaktične in semantične natančnosti ter popolnosti v obeh preglednicah, tj. preglednici »Starsi« in preglednici »Otroci«. Z vidika obeh preglednic pa stanje še zmeraj ni zadovoljivo. V preglednico »Otroci« namreč lahko dodamo posameznega otroka, ki ni vezan na starše. Navedeno pomanjkljivost odpravim s postavitvijo odnosa med obema preglednicama, tj. tujim ključem.

```
ALTER TABLE Otroci
ADD CONSTRAINT FK_ZaporednaStevilkaStarsa FOREIGN KEY (ZaporednaStevilka)
REFERENCES Starsi (ZaporednaStevilka)
```

SLIKA 22: PODATKOVNI MODEL, MANJŠA DOSLEDNOST – SCENARIJ 1



V preglednico »Otroci« po izvedenih spremembah ne moremo vnesti otroka, če prej ne vnesemo enega od njegovih staršev. Prav tako ne moremo izbrisati staršev iz preglednice »Starsi«, ki ima pripadajoč zapis v preglednici »Otroci«. S tem sem zagotovil že določeno doslednost, tj. usklajenost s semantičnim pravilom, da v vrtec sprejemamo otroke znanih staršev.

Kljub znatnemu napredku v kakovosti modela še zmeraj ni izvedenega pravila, da v vrtec sprejemamo največ tri otroke posameznih staršev. Omenjeno pravilo po navadi uvedemo v programski kodi z različni kontrolami oziroma preverjanji. Vendar bom zaradi proučevanja vpliva FIRPM na kakovost podatkov omenjeno pravilo uvedel z različnim modeliranjem.

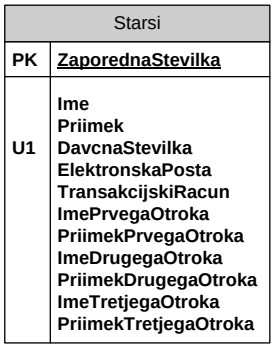
Na sliki 22 imamo model, ki je popolnoma normaliziran, tj. brez ponavljajočih se vrednosti, delnih in tranzitivnih odvisnosti, vendar omogoča vnos N otrok posameznih staršev, jaz pa želimo omejiti število otrok, zato moram izvesti naslednjo programsko kodo:

```
CREATE TABLE Starsi (
 ZaporednaStevilka     INT PRIMARY KEY,
 Ime                   VARCHAR(150) NOT NULL,
 Priimek               VARCHAR(150) NOT NULL,
 DavcnaStevilka        VARCHAR(150) NOT NULL,
 ElektronskaPosta      VARCHAR(100) NOT NULL,
 TransakcijskiRacun    VARCHAR(150) NOT NULL,
 ImePrvegaOtroka       VARCHAR(150) NOT NULL,
 PriimekPrvegaOtroka   VARCHAR(150) NOT NULL,
 ImeDrugegaOtroka      VARCHAR(150) NOT NULL DEFAULT '',
 PriimekDrugegaOtroka  VARCHAR(150) NOT NULL DEFAULT '',
 ImeTretjegaOtroka     VARCHAR(150) NOT NULL DEFAULT '',
 PriimekTretjegaOtroka VARCHAR(150) NOT NULL DEFAULT '',)
GO
ALTER TABLE Starsi
ADD CONSTRAINT Ime_NI_Neznano CHECK (Ime<>''),
CONSTRAINT Priimek_NI_Neznan CHECK (Priimek<>''),
CONSTRAINT Imeotroka_NI_Neznano CHECK (ImePrvegaOtroka<>''),
CONSTRAINT PriimekOtroka_NI_Neznan CHECK (PriimekPrvegaOtroka<>''),
CONSTRAINT TRR_Kontrola  CHECK (dbo.KontrolaTRR(TransakcijskiRacun)=1),
CONSTRAINT Davcna_Edinstvenost  UNIQUE NONCLUSTERED (DavcnaStevilka)
```

Namesto dveh preglednic obstaja samo ena, v kateri so polja za vnos nazivov otrok. Obveznost vnosa pri nazivih otrok sem postavil samo na polju za vnos prvega otroka. S tem zahtevam vnos najmanj enega otroka, kar je logično iz navedenega primera, saj je brezpredmetno vpisovati podatke staršev, ki nimajo vpisanega nobenega otroka. Po drugi strani spremenjeni model omogoča vnos največ treh otrok, kar je tudi semantično pravilo vrtca, in s tem model zagotavlja doslednost. Če bi obveznost vnosa postavil tudi na polji za vnos naziva za drugega oziroma tretjega otroka, bi postavili drugačno semantično pravilo. To pravilo bi se glasilo: v vrtec sprejemamo otroke znanih staršev, in sicer najmanj dva in največ tri.

Podatkovni model na sliki 23 je popolnoma normaliziran, prav tako kot model na sliki 22. Z omenjenim scenarijem sem prikazal, kako lahko z dobrim FIRPM zagotavljamo večjo kakovost podatkov.

SLIKA 23: PODATKOVNI MODEL, VEČJA DOSLEDNOST – SCENARIJ 1



Za preverjanje podatkovnega modela bom izvedel naslednje ukaze:

```
INSERT INTO Starsi
(ZaporednaStevilka,Ime,Priimek,DavcnaStevilka,ElektronskaPosta,TransakcijskiRa
cun,ImePrvegaOtroka,PriimekPrvegaOtroka)
  VALUES (1,'Janez','Novak','23414867','janez.novak@gmail.com','04881-
0000921','Anže','Novak')
GO
INSERT INTO Starsi
(ZaporednaStevilka,Ime,Priimek,DavcnaStevilka,ElektronskaPosta,TransakcijskiRa
cun,ImePrvegaOtroka,PriimekPrvegaOtroka)
  VALUES (2,'Andrej','Meglic','23414867','andrej.meglic@gmail.com','04881-
03400921','Luka','Meglič')
GO
INSERT INTO Starsi
(ZaporednaStevilka,Ime,Priimek,DavcnaStevilka,ElektronskaPosta,TransakcijskiRa
cun,ImePrvegaOtroka,PriimekPrvegaOtroka)
  VALUES (3,'Andreja','Dolinar','73415867','andreja.dolinar_gmail.com','04881-
000','Nina','Dolinar')
```

Prvi ukaz se uspešno izvede, drugi in tretji pa neuspešno. V drugem ukazu sta dve napaki, kršitev deklarativne omejitve polja »DavcnaStevilka« in polja »TransakcijskiRacun«, v tretjem ukazu pa deklarativna omejitev »ElektronskaPosta« in polja »TransakcijskiRacun«. Iz izvedenega modela oziroma simulacije je mogoče sklepati na naslednje:

- Normalizacija ima možnost, da zagotavlja doslednost, in sicer zmanjšuje število ponavljanj, hkrati pa s pravilnim modeliranjem omogočamo izvedbo poslovnih pravil.
- Deklarativne omejitve lahko zagotovijo visoko raven sintaktične in semantične natančnosti, ne morejo pa zagotoviti popolne natančnosti. Zagotovijo lahko tudi visoko raven popolnosti.

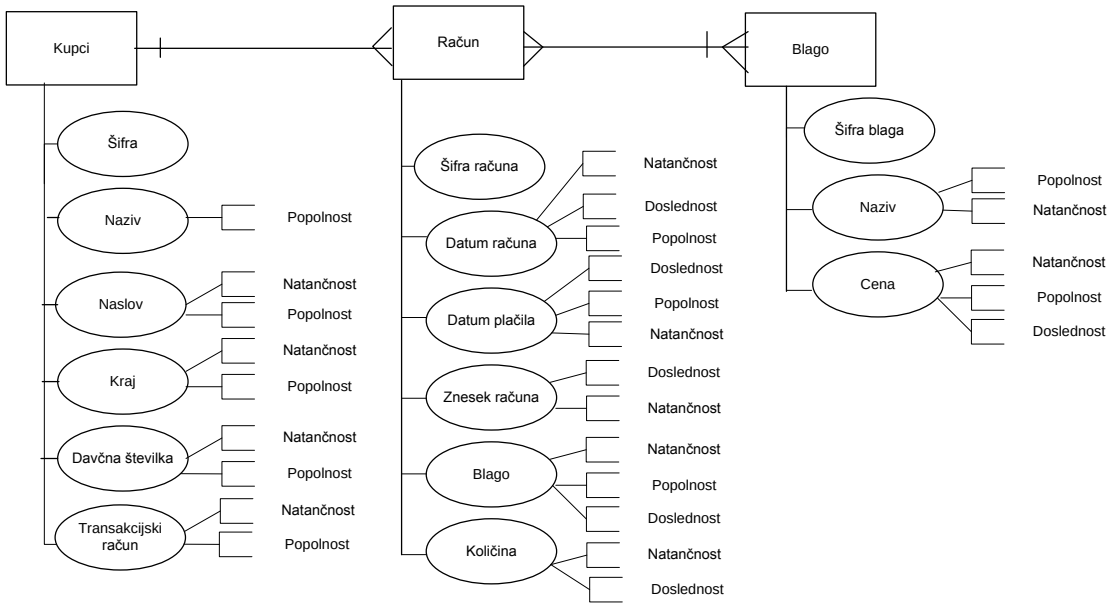
- Podatkovni tipi, še posebno zapleteni, so možno sredstvo za doseganje sintaktične in semantične natančnosti.
- Z glavnimi in tujimi ključi lahko pripomoremo k večji doslednosti.

9.3.2 Načrtovanje podatkovnega modela – scenarij 2: Prodaja blaga

Scenarij: Združba prodaja blago kupcem, za katere izdaja račun. Mnogokrat se zgodi, da kupci spet pokličejo združbo za izstavitev kopije računa. Račun mora biti enak kot izvirnik, neodvisno od sprememb podatkov, ki so se morda zgodile v vmesnem času, na primer preselitev kupca, zvišanje cene.

Kot pri prvem scenariju tudi v drugem najprej ustvarim razširjen entitetno-relacijski diagram z vidikom kakovosti (slika 24).

SLIKA 24: RAZŠIRJENI ENTITETNO-RELACIJSKI DIAGRAM Z VIDIKOM KAKOVOSTI – SCENARIJ 2



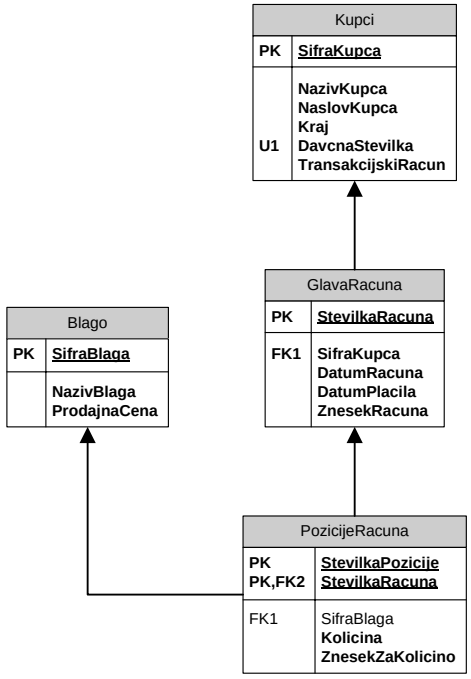
Z naslednjo programsko kodo ustvarim fizični podatkovni model na sliki 25:

```
CREATE TABLE Kupci (
SifraKupca          BIGINT PRIMARY KEY,
NazivKupca          VARCHAR(200) NOT NULL DEFAULT '',
NaslovKupca         VARCHAR(500) NOT NULL DEFAULT '',
Kraj                VARCHAR(100) NOT NULL DEFAULT '',
DavcnaStevilka      VARCHAR(15)  NOT NULL DEFAULT '',
TransakcijskiRacun VARCHAR(16)  NOT NULL DEFAULT '')
GO
ALTER TABLE Kupci ADD
CONSTRAINT TRR_Kontrola  CHECK (dbo.KontrolaTRR(TransakcijskiRacun)=1),
CONSTRAINT DavcnaSt_Edinstvenost  UNIQUE NONCLUSTERED (DavcnaStevilka),
```

```
CONSTRAINT NazivKupca_NI_PRAZEN  CHECK (NazivKupca<>''),
CONSTRAINT NaslovKupca_NI_PRAZEN CHECK (NaslovKupca<>''),
CONSTRAINT KrajKupca_NI_PRAZEN CHECK (Kraj<>'')
GO
CREATE TABLE Blago(
SifraBlaga       CHAR(10) PRIMARY KEY,
NazivBlaga         VARCHAR(200) NOT NULL DEFAULT '',
ProdajnaCena       DECIMAL(12,6)NOT NULL DEFAULT 0)
GO
ALTER TABLE Blago ADD
CONSTRAINT NazivBlaga_NI_PRAZEN CHECK (NazivBlaga<>''),
CONSTRAINT ProdajnaCena_JE_POZITIVNA  CHECK (ProdajnaCena>0)
GO
CREATE TABLE GlavaRacuna(
StevilkaRacuna     BIGINT PRIMARY KEY,
SifraKupca         BIGINT NOT NULL DEFAULT '',
DatumRacuna        DATETIME NOT NULL DEFAULT GETDATE(),
DatumPlacila       DATETIME NOT NULL DEFAULT GETDATE(),
ZnesekRacuna       DECIMAL(19,2) NOT NULL DEFAULT 0)
GO
ALTER TABLE GlavaRacuna ADD
CONSTRAINT SifraKupca_TUJ_KLJUC FOREIGN KEY
  (SifraKupca) REFERENCES Kupci(SifraKupca),
CONSTRAINT DatumPlacila_JE_VECJI CHECK (DatumPlacila>=DatumRacuna),
CONSTRAINT ZnesekRacuna_POZITIVEN CHECK (ZnesekRacuna>0)
GO
CREATE TABLE PozicijeRacuna(
StevilkaPozicije   BIGINT NOT NULL,
StevilkaRacuna     BIGINT NOT NULL,
SifraBlaga         CHAR(10) DEFAULT '',
Kolicina           DECIMAL(9,2) NOT NULL DEFAULT 0,
ZnesekZaKolicino   DECIMAL(19,2) NOT NULL DEFAULT 0)
GO
ALTER TABLE PozicijeRacuna ADD
CONSTRAINT PK_POZ_RAC PRIMARY KEY (StevilkaPozicije,StevilkaRacuna),
CONSTRAINT StevilkaRacuna_TUJ_KLJUC FOREIGN KEY
  (StevilkaRacuna) REFERENCES GlavaRacuna(StevilkaRacuna),
CONSTRAINT SifraBlaga_TUJ_KLJUC FOREIGN KEY
  (SifraBlaga) REFERENCES Blago(SifraBlaga),
CONSTRAINT Kolicina_POZITIVEN CHECK (Kolicina>0),
CONSTRAINT ZnesekZaKolicino_POZITIVEN CHECK (ZnesekZaKolicino>0)
```

Če pregledam ustvarjeni model, vidim, da je model popolnoma normaliziran. V modelu ne obstajajo ponavljajoče se sestavine, delne in tranzitivne odvisnosti. Uvedeni so ključi, ustrezni podatkovni tipi ter deklarativne omejitve. Z navedenimi funkcijami bi moral doseči visoko stopnjo sintaktične in semantične natančnosti, kot tudi popolnosti in doslednosti. Pa je res tako?

SLIKA 25: ZAČETNI PODATKOVNI MODEL – SCENARIJ 2



Model je zadovoljiv z vidika natančnosti in popolnosti, pri razsežnosti doslednost pa ima precej pomanjkljivosti in potrebuje izboljšave. Težave lahko pričakujem na dveh področjih, in sicer pri nedoslednosti pri vnosu kraja kupca, kar lahko oteži poznejše načrte glede analize prodaje po krajih, težave pa so tudi pri izpolnjevanju zahteve o istovetnosti računa skozi čas. Prodajna cena za posamezno blago obstaja v preglednici »Blago«, naslov kupca ter kraj pa v preglednici »Kupci«. Kaj se torej zgodi, če se od datuma izdaje računa do zahtevka po kopiji računa spremenita cena blaga ali naslov kupca ali oboje?

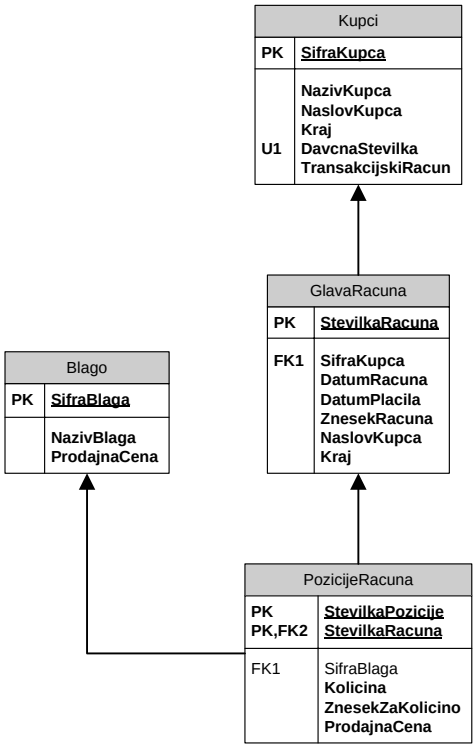
Ustvarjeni model v tem primeru ne zagotovi doslednosti, saj bi kupec sprejel račun z drugačnimi vrednostmi v primerjavi s pristnim računom. Za razrešitev težav glede doslednosti obstaja več možnosti, prikazal pa bom uporabo denormalizacije.

Z uporabo denormalizacije dodam polje »ProdajnaCena« iz preglednice »Blago« v preglednico »PozicijeRacuna«, v preglednico »GlavaRacuna« pa dodam polji »Naslov« in »Kraj«.

```
ALTER TABLE PozicijeRacuna
 ADD ProdajnaCena DECIMAL(12,6)NOT NULL DEFAULT 0,
 CONSTRAINT ProdajnaCena_JE_POZITIVNA_1  CHECK (ProdajnaCena>0)
GO
ALTER TABLE GlavaRacuna
ADD NaslovKupca VARCHAR(500) NOT NULL DEFAULT '',
Kraj VARCHAR(100) NOT NULL DEFAULT '',
CONSTRAINT NaslovKupca_NI_PRAZEN_1 CHECK (NaslovKupca<>''),
CONSTRAINT KrajKupca_NI_PRAZEN_1 CHECK (Kraj<>'')
```

Po denormalizaciji dobim model na sliki 26.

SLIKA 26: PODATKOVNI MODEL PO DENORMALIZACIJI – SCENARIJ 2



Pri vnosu računa prepišem v preglednici »GlavaRacuna« in »PozicijeRacuna« še dodatne podatke ter s tem zagotovim doslednost. Zaradi spremenjenega modela pri izpisu računa ne potrebujem več vseh štirih preglednic, temveč samo še dve. Poleg tega lahko močno povečam natančnost polja »ZnesekZaKolicino«, če ga opredelim kot izvedeno polje. Možnost izvedenih polj (angl. *derived columns*) je spet odvisna od funkcij posameznega SUBP.

```
CREATE TABLE PozicijeRacuna(
StevilkaPozicije   BIGINT NOT NULL,
StevilkaRacuna     BIGINT NOT NULL,
SifraBlaga       CHAR(10) NOT NULL DEFAULT '',
Kolicina           DECIMAL(9,2) NOT NULL DEFAULT 0,
ProdajnaCena       DECIMAL(12,6)NOT NULL DEFAULT 0,
ZnesekZaKolicino AS Kolicina*ProdajnaCena)
```

Z izvedenim poljem sem zmanjšal potrebo po nadzoru kakovosti podatkov tega polja, saj zagotovitev kakovosti polj, ki sestavljajo izvedeno polje, neposredno zagotavlja tudi kakovost izvedenega stolpca.

Z »denormalizacijo« sem dosegel, da model zagotavlja doslednost. Pravzaprav ne gre za denormalizacijo, ker je naslov dejansko lastnost računa. Praviloma bi moral časovno razsežnost izvesti drugače, vendar je pravilna izvedba časovne razsežnosti dokaj zapleteno opravila in je

zato po navadi smotrno uporabljati kompromis ter slediti primerom najboljše prakse. Še boljše bi bilo obstoječi model razširiti in uporabiti zapletene podatkovne tipe, kamor bi shranil račun v obliki pdf ali XML. Če bi se odločil za pdf, bi moral obdržati preglednici »GlavaRacuna« in »PozicijeRacuna«, saj poizvedbe SQL v listinah pdf niso mogoče. Če bi se odločil za uporabo podatkovnega tipa XML, postanejo določena polja odvečna.

Naslednja programska koda ustvari preglednico, ki vsebuje zapleteni podatkovni tip XML:

```
CREATE TABLE GlavaRacuna_XML (
StevilkaRacuna BIGINT NOT NULL,
SifraKupca     BIGINT NOT NULL,
Racun          XML)
GO
ALTER TABLE GlavaRacuna_XML
 ADD CONSTRAINT PK_GlavaRacuna_XML PRIMARY KEY (StevilkaRacuna)
```

V polje »Racun« lahko vnesem celoten račun, kar storim s spodnjim ukazom. Tako sem ugodil pravilu o doslednosti podatkov na računu.

```
INSERT INTO GlavaRacuna_XML
SELECT 1,100,
'(<Racun>
   <DatumNarocila>20080110</DatumNarocila>
   <DatumPlacila>20080118</DatumPlacila>
   <Naslov>Koroška cesta 76</Naslov>
   <Kraj>Tržič</Kraj>
   <ZnesekRacuna>1960</ZnesekRacuna>
  <PozicijeRacuna>
    <PozicijaRacuna1>
            <SifraBlaga>010023</SifraBlaga>
            <NazivBlaga>Hladnilnik Gorenje MDF 1000</NazivBlaga>
            <Kolicina>1</Kolicina>
            <ProdajnaCena>1345</ProdajnaCena>
            <SkupajZaKolicino>1340</SkupajZaKolicino>
    </PozicijaRacuna1>
    <PozicijaRacuna2>
            <SifraBlaga>010026</SifraArtikla>
            <NazivBlaga>Pečica Gorenje PET 65</NazivBlaga>
            <Kolicina>1</Kolicina>
            <ProdajnaCena>620</ProdajnaCena>
            <SkupajZaKolicino>620</SkupajZaKolicino>
    </PozicijaRacuna2>
  </PozicijeRacuna>
</Racun>'
UNION ALL
SELECT 2,100,
'(<Racun>
   <DatumNarocila>20080120</DatumNarocila>
   <DatumPlacila>20080128</DatumPlacila>
   <Naslov>Koroška cesta 76</Naslov>
   <Kraj>Tržič</Kraj>
   <ZnesekRacuna>1000</ZnesekRacuna>
  <PozicijeRacuna>
```

```
    <PozicijaRacuna1>
            <SifraBlaga>010024</SifraBlaga>
            <NazivBlaga>Miza Sonja Lip Bled</NazivBlaga>
            <Kolicina>1</Kolicina>
            <ProdajnaCena>1000</ProdajnaCena>
            <SkupajZaKolicino>1000</SkupajZaKolicino>
    </PozicijaRacuna1>
  </PozicijeRacuna>
</Racun>'
UNION ALL
SELECT 3,101,
'(<Racun>
   <DatumNarocila>20080120</DatumNarocila>
   <DatumPlacila>20080128</DatumPlacila>
   <Naslov>Cesta na Brdo 3</Naslov>
   <Kraj>Kranj</Kraj>
   <ZnesekRacuna>500</ZnesekRacuna>
  <PozicijeRacuna>
    <PozicijaRacuna1>
            <SifraBlaga>010024</SifraBlaga>
            <NazivBlaga>Mikrovalovna pečica Candy IX 300</NazivBlaga>
            <Kolicina>1</Kolicina>
            <ProdajnaCena>500</ProdajnaCena>
            <SkupajZaKolicino>500</SkupajZaKolicino>
    </PozicijaRacuna1>
  </PozicijeRacuna>
</Racun>'
```

Podatkovni tip XML omogoča, da je podatkovni model ustrezen, da zagotavlja visoko kakovost podatkov. V SQL 2005 za poizvedbe po podatkovnih tipih XML uporabljamo XQUERY. Naslednji XQUERY vrne primerjavo med starim in novim naslovom, kjer se naslova med seboj razlikujeta:

```
SELECT Racun.value('(/Racun/Naslov)[1]', 'varchar(max)')+' '+
       Racun.value('(/Racun/Kraj)[1]', 'varchar(max)') AS StarNaslov,
       Kupci.NaslovKupca+' '+Kupci.Kraj AS NovNaslov
  FROM GlavaRacuna_XML
   JOIN Kupci ON GlavaRacuna_XML.SifraKupca=Kupci.SifraKupca
    WHERE Racun.value('(/Racun/Naslov)[1]', 'varchar(max)')<>Kupci.NaslovKupca
```

V prvem scenariju sem prikazal, da deklarativne omejitve lahko vplivajo na semantično natančnost v okviru posamezne preglednice oziroma natančneje v okviru posameznega polja v preglednici. V nadaljevanju pa bom prikazal, kako lahko deklarativne omejitve vplivajo tudi na usklajenost s semantičnimi pravili v okviru več preglednic, tj. v celotni zbirki podatkov. Drugi scenarij bom dopolnil s poslovnim pravilom,[14] da kupcem z boniteto A priznam 25-odstotni popust, z boniteto B 10-odstotni popust, ostalim pa dodatnega popusta ne priznam. Takšno pravilo je po navadi izvedeno v poslovni logiki na srednji ravni. Toda če je to pravilo ključno za poslovanje združbe, ga je treba izvesti »čim bližje« podatkovnemu tipu. Tako bi bilo morda

[14] Poslovna pravila so podmnožica semantičnih pravil, zato usklajenost s poslovnimi pravili pomeni tudi usklajenost s semantičnimi pravili.

smiselno izvesti zelo pogosto poslovno pravilo, da kupcem, katerih dolg preseže določen znesek, onemogočimo dobavo blaga oziroma izvedbo storitve.

Za potrebo izvajanja poslovnega pravila glede popustov dodam dve novi polji, eno v preglednico »Kupci«, drugo v preglednico »PozicijeRacuna«.

```
ALTER TABLE Kupci
 ADD Boniteta Char(1) NOT NULL DEFAULT ''
GO
ALTER TABLE PozicijeRacuna
  ADD Popust DECIMAL(5,2) NOT NULL DEFAULT 0
```

Poleg tega vnovič opredelim izvedeno polje »ZnesekZaKolicino«, saj moram upoštevati tudi popust:

```
ALTER TABLE PozicijeRacuna
  DROP COLUMN ZnesekZaKolicino
GO
ALTER TABLE PozicijeRacuna
 ADD  ZnesekZaKolicino  AS  Kolicina*ProdajnaCena*(CASE  WHEN  Popust>0  THEN
Popust/100 ELSE 1 END)
```

Do uskladitve podatkovnega modela s poslovnim pravilom manjka le še deklarativna omejitev na polju »Popust«. Spet bom prikazal primer UDF (*User defined function*), enako funkcijo pa bi lahko izvedel tudi z zapletenim podatkovnim tipom. Programska koda, ki ustvari funkcijo, je:

```
CREATE FUNCTION [dbo].[PreverbaPopusta](@StevilkaRacuna BIGINT,@Popust
DECIMAL(5,2))
RETURNS INT AS
BEGIN
   DECLARE @Boniteta CHAR(1)
   DECLARE @Napaka int
SET @Napaka=1

SELECT @Boniteta=K.Boniteta
  FROM Kupci AS K
   JOIN GlavaRacuna AS G ON G.SifraKupca=K.SifraKupca
      WHERE G.StevilkaRacuna=@StevilkaRacuna

IF @Boniteta='A' AND @Popust<>25
     SET @Napaka=-1
ELSE IF @Boniteta='B' AND @Popust<>10
     SET @Napaka=-1
ELSE IF @Boniteta='' AND @Popust<>0
     SET @Napaka=-1

RETURN @Napaka
END
```

V polje »Popust« v preglednici »PozicijeRacuna« dodam deklarativno omejitev in podatkovni model je usklajen s poslovnim pravilom, kar zagotovi natančnost.

```
ALTER TABLE PozicijeRacuna
ADD CONSTRAINT Preverba_Popusta CHECK
(dbo.PreverbaPopusta(StevilkaRacuna,Popust)=1)
```

Prikazani scenarij pokaže, da normalizacija vpliva na večjo doslednost podatkov. Na večjo doslednost vplivajo tudi deklarativne omejitve. Večjo doslednost lahko dosežemo tudi z uporabo pravilnih podatkovnih tipov. Normalizacija pa ima še en vpliv, in sicer zagotavlja večjo popolnost podatkov. Če bi enak podatek morali vpisovati na več mestih, bi se hitro lahko primerilo, da bi določen podatek pozabili vnesti na določenem mestu in v zbirki podatkov bi imeli večje število manjkajočih vrednosti, poleg tega pa je normalizacija tudi postopek, v katerem opredelimo entitete in jih logično oblikujemo.

9.3.3 Razprava na podlagi modeliranja

Z modeliranjem sem prikazal, kako lahko uporaba pravil, ki jih omogoča relacijski podatkovni model ter sodobni SUBP, vpliva na večjo kakovost podatkov. V preglednici 16 so navedene lastnosti FIRPM ter razsežnosti kakovosti podatkov, na katere lahko vplivajo posamezne lastnosti FIRPM. Najmanj očitna je vloga glavnih ključev, ki sicer ne dovoljujejo manjkajočih vrednosti ter ponavljanja enakih podatkov (vpliv na popolnost in doslednost), vendar pa je ta vloga bolj tehnične kot vsebinske narave. Podatki, ki tvorijo glavni ključ, so po navadi suhoparni in nimajo velike vloge za uporabnike. To je še posebno izrazito pri umetnih glavnih ključih, ko njihovo vlogo po navadi opravlja polje, ki hrani zaporedno številko zapisa. Takšen način opredeljevanja glavnega ključa izhaja iz želje razvijalcev podatkovnega modela po čim hitrejšem izvajanju poizvedb, saj vsako dodatno polje v glavnem ključu upočasni izvajanje poizvedb, hkrati pa poslabša tudi preglednost podatkovnega modela. Kandidati za glavni ključ ne morejo biti »vsebinsko močnejša« polja, saj njihova vrednost v preglednici po navadi ni edinstvena.

Znesek računa, znesek plače, regija, cena blaga, vrsta avtomobila, zakonski stan, lastništvo hiše itn. niso edinstveni znotraj preglednice oziroma so edinstveni v zelo redkih primerih in posledično glavni ključ, sestavljen iz teh polj, ni ustrezen. Vloga glavnega ključa je predvsem v zagotavljanju spoštovanja semantičnih pravil, torej v zagotavljanju doslednosti, na primer en zapis študenta v šifrantu študentov, en zapis kupca v šifrantu kupcev, en zapis blaga oziroma storitve v ustreznem šifrantu, ter v zagotavljanju popolnosti, na primer obvezen vnos EMŠO, davčne številke, če so ta polja opredeljena kot glavni ključ.

Če hočemo zagotoviti doslednost ter popolnost v navedenih primerih, moramo ustrezno izbrati polja, ki sestavljajo glavni ključ. Pri študentih in kupcih bi bilo najboljše uporabiti »naravni« glavni ključ, o katerem sem že razpravljal. Vendar takšna odločitev ni vedno mogoča, saj smo postali odprta družba in na naše fakultete se vpisujejo tudi študenti iz tujine, ki na primer nimajo

davčne ali matične številke, slovenske združbe poslujejo z združbami iz tujine, ki spet nimajo enakih lastnosti, zato smo velikokrat prisiljeni uporabiti »umetni« glavni ključ.

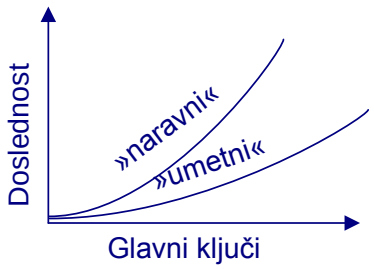
PREGLEDNICA 16: PROUČEVANJE VPLIVA FIRPM NA KAKOVOST PODATKOV S POMOČJO MODELIRANJA

Lastnost FIRPM	Lahko vpliva na razsežnosti kakovosti podatkov
Normalizacija	Doslednost, popolnost
Podatkovni tipi	Natančnost, doslednost
Deklarativne omejitve	Natančnost, popolnost, doslednost
Tuji ključi	Doslednost
Glavni ključi	Doslednost, popolnost

»Umetni« glavni ključ po navadi ne vsebuje lastnosti, ki bi preprečevale vnovičen vnos že vnesenega zapisa. Vpisna številka študenta, šifra kupca, šifra blaga so lastnosti, ki so omenjenim entitetam »umetno« dodane in z njimi niso neposredno povezane. Na podlagi povedanega lahko sklepam, da »naravni« glavni ključi v primerjavi z »umetnimi« glavnimi ključi bolje zagotavljajo doslednost (slika 27), vendar je vpliv najbrž zanemarljiv. Glavni ključi pa preprečujejo vnos več praznih vrednosti oziroma vrednosti NULL in posledično lahko sklepam, da prispevajo k večji popolnosti podatkov.

Vendar pa so tudi »umetni« ključi ustrezni, če jih dopolnimo z deklarativnimi omejitvami. V podatkovnem modelu bi kot glavni ključ izbrali zaporedno številko na primer zaposlenega, z deklarativno omejitvijo pa bi določili edinstvenost polja EMSO.

SLIKA 27: PRIKAZ ODVISNOSTI MED GLAVNIMI KLJUČI IN DOSLEDNOSTJO



Enako, kot je bilo povedano in prikazano za glavne ključe, velja tudi za tuje ključe, le da slednji pravzaprav nimajo vpliva na popolnost. Če imamo v glavni preglednici tudi eno prazno vrednost v stolpcu, ki predstavlja glavni ključ, potem lahko v odvisni preglednici v povezan stolpec vnesemo poljubno število praznih vrednosti.

Z obravnavanimi scenariji sem nakazal, kako lahko tuji ključi vplivajo na usklajenost s semantičnimi pravili. Njihova vloga je v tem pogledu neprecenljiva in pomembno prispeva k

celovitosti podatkov. Tuji ključi preprečujejo brisanje zapisov, če imajo ti zapisi tudi odvisne zapise, hkrati pa preprečujejo vnos odvisnih zapisov, če slednji nimajo najprej vnesenega osnovnega zapisa.

To omogoča celovitost podatkov v SSOP in manjšo možnost za napake v skladiščih podatkov ter posledično bolj pravilne analize. Brez tujih ključev bi verjetno imeli precej več napak v podatkih. Vrstice računa, na primer, ne bi imele pripadajoče glave računa, kjer imamo navedenega kupca in različne datume. Prav tako bi se vrstice računa sklicevale na poslovne učinke, ki v zbirki podatkov ne obstajajo, in posledično bi v združbi imeli napačno predstavo o prodaji, medtem ko bi bile zapletenejše analize popolnoma nekoristne.

Podoben vpliv na kakovost podatkov kot tuji ključi ima tudi proces normalizacije. S procesom normalizacije zagotovimo manjše število ponavljanj podatkov, za katerega sta Rudra in Yeo (1999) ugotovila, da povečuje doslednost. Vendar zmanjševanje števila ponavljanj podatkov ni edini način, kako normalizacija povečuje doslednost. Z modeliranjem sem pokazal, da način izvedbe normalizacije uveljavlja pravila in zmanjša potrebo po drugih principih relacijskega modela. Primer vrtca je pokazal, kako lahko popolnoma normaliziran podatkovni model da drugačne poslovne rešitve. Drugi primer pa je pokazal, kako lahko nasprotni proces, tj. denormalizacija, prav tako zagotavlja doslednost, a hkrati s tem zmanjšuje popolnost. Uporaba normalizacije mora biti premišljena, da zagotovimo ustrezen podatkovni model, ki bo uveljavljal poslovna pravila, predvsem tista, ki vsebujejo časovno razsežnost, hkrati pa ne bo dovoljeval nepotrebnega ponavljanja podatkov. Z ustrezno denormalizacijo lahko zagotovimo upoštevanje časovne razsežnosti, kar je tudi odgovor kritikom relacijskega podatkovnega modela, ki mu očitajo slabšo podporo časovni razsežnosti.

Za razpravo sta ostali še dve lastnosti, in sicer deklarativne omejitve ter podatkovni tipi. Za deklarativne omejitve sem pokazal, kako vplivajo na oba vidika razsežnosti »natančnost«, na razsežnost »popolnost«, in sicer tiste popolnosti, ki jo Pipino, Lee in Wang (2002) enačijo s stolpčno integriteto, ter na razsežnost »doslednost«. Z modeliranjem sem vnovič nakazal vpliv deklarativnih omejitev na doslednost podatkov, čeprav sta to storila že Rudra in Yeo (1999). Njuno trditev pa sem v nadaljevanju še dodatno preveril z obdelavo podatkov, ki sem jih zbral z anketo. Uporaba deklarativnih omejitev zahteva tehten premislek, saj njihova prevelika uporaba povečuje zapletenost podatkovnega modela in posledično otežuje vzdrževanje. Njihova uporaba je smotrna na poljih, ki hranijo pomembne podatke za poslovanje združbe, ter pri uveljavljanju ključnih semantičnih pravil.

Podobno je tudi s podatkovnimi tipi, saj morajo biti ustrezni glede na tip podatka, ki ga želimo hraniti v polju. Izogibati se je treba znakovnim podatkovnim tipom za polja, v katerih bomo

shranjevali samo številčne podatke, saj s tem povečamo možnost za sintaktično nenatančne podatke, hkrati pa omejimo nabor operacij, ki jih nad podatki lahko izvajamo. Nesmotrna in odsvetovana je tudi uporaba znakovnih podatkovnih tipov za datume, saj imajo datumska polja nabor posebnih operacij (na primer v SQL 2005 *month()*, *datediff()*, *datepart()*). Podatkovni tipi morajo ustrezati nameravani uporabi. Z novejšim SUBP se je pojavila tudi možnost izdelave lastnih, bolj zapletenih podatkovnih tipov. Morda se zdi privlačno uporabiti čim večje število lastnih podatkovnih tipov, vendar ob podrobnejšem premisleku navdušenje hitro usahne. V prid razširjeni uporabi lastnih podatkovnih tipov sicer govori predvsem dejstvo, da lahko učinkovito zagotovimo natančnost ter doslednost podatkov. Z izdelavo lastnih podatkovnih tipov bi lahko preprečili tudi vnos praznih vrednosti oziroma vrednosti NULL, vendar je to z vidika učinkovitosti popolnoma neupravičeno, zato podatkovnih tipov kot sredstva za doseganje popolnosti ne bom vključil v model.

Za neuporabo lastnih podatkovnih tipov lahko naštejem kar nekaj dejstev. Velika ovira so težavno vzdrževanje, slabša razumljivost, ki lahko zelo omeji uporabo lastnega podatkovnega tipa, ter stroški izdelave. Bolj zapleten je lastni podatkovni tip, več energije moramo vložiti v njegovo izdelavo in vzdrževanje. Tako kot pri deklarativnih omejitvah je uporaba lastnih podatkovnih tipov smiselna pri ključnih podatkih in poslovnih pravilih. Med deklarativnimi omejitvami in podatkovnimi tipi je kar nekaj podobnosti, a hkrati tudi razlik v skladu z namenom uporabe. Pomembna razlika je v tem, da s podatkovnim tipom določimo tudi operacije nad podatki v polju, medtem ko deklarativne omejitve tega ne določajo. SUBP deluje tako, da najprej preveri ustreznost podatkovnega tipa, sledi preverjanje privzetih vrednosti (angl. *default values*) ter vrednosti NULL/NOT NULL in šele nato se preverja usklajenost podatka z deklarativnimi omejitvami. Poudariti je treba še eno razliko med lastnimi podatkovnimi tipi ter deklarativnimi omejitvami, in sicer SUBP omogoča »izklop/onemogočanje« preverjanja podatka z deklarativnimi omejitvami (na primer v SQL 2005 bi to storili z naslednjim ukazom: `ALTER TABLE Kupci WITH NOCHECK CHECK CONSTRAINT ALL`), medtem ko podatkovnega tipa ne moremo onemogočiti. Menim, da je bolje uporabiti deklarativne omejitve, seveda tam, kjer je to mogoče. K temu me napeljuje naslednji razmislek. Načrtovalci zbirk podatkov znajo bolje uporabljati jezik SQL, ker je ta na trgu že več kot 20 let, medtem ko se jeziki CLR (*common languages runtime*) v povezavi z zbirkami podatkov uporabljajo le nekaj let. Jeziki CLR so bili sprva namenjeni razvoju programske kode srednje in odjemalske ravni, šele v zadnjem času pa se je njihova uporabnost razširila še na izdelavo lastnih podatkovnih tipov, najbrž tudi zaradi želje velikih razvijalcev SUBP, da bi razvoj podatkovnega modela približali tudi do zdaj ozko usmerjenim programerjem. Slednji v večini primerov niso vešči pri načrtovanju zbirk podatkov, čeprav je res, da programerji po navadi opravljajo več nalog. Poleg tega ima veliko uporabniških rešitev korenine v preteklosti, ko

še ni bilo mogoče uporabljati lastnih podatkovnih tipov, in bi sedanja prevelika sprememba že obstoječe uporabniške rešitve verjetno bila stroškovno neupravičena.

Moj nasvet po proučevanju razmerja med lastnimi podatkovnimi tipi in deklarativnimi omejitvami je v prid uporabi slednjih. Če torej določeno težavo lahko učinkovito rešimo z uporabo deklarativnih omejitev ter lastnimi podatkovnimi tipi, raje izberimo deklarativne omejitve. Če pa težave lahko odpravimo samo z lastnimi podatkovnimi tipi, izbire nimamo. Vendar menim, da takšnih težav ni veliko, vsaj na področju PIS, bolj običajne so v ozko usmerjenih informacijskih sistemih, ki so namenjeni reševanju posebnih težav, kot na primer športni rezultati na televiziji ORF.

Z modeliranjem sem izpeljal model, ki predpostavlja:

$$Q = f(K_{2j}) = g(D_{2j}, U_{2j})$$

$$D_{2j} = k(P_n)$$

$D_{21} = k(P_1, P_2)$
$D_{22} = k(P_1, P_2, P_3, P_4, P_5)$
$D_{23} = k(P_2, P_3, P_4)$,

kjer je D_{21} – natančnost, D_{22} – doslednost, D_{23} – popolnost, P_1 – podatkovni tipi, P_2 – deklarativne omejitve, P_3 – normalizacija, P_4 – glavni ključi, P_5 – tuji ključi.

Model tudi predpostavlja, da je najtežje vplivati na razsežnost natačnost, veliko lažje pa na razsežnosti popolnost in doslednost. Modeliranje nakazuje, da je povezava med neodvisnimi spremenljivkami in odvisnimi spremenljivkami pozitivna.

Iz modela izhajajo naslednje povezave:

- Ustreznejši podatkovni tipi ter več deklarativnih omejitev naj bi izboljšalo natančnost podatkov.
- Ustreznejši podatkovni tipi, več deklarativnih omejitev, dosledna uporaba glavnih in tujih ključev ter spoštovanje normalizacije naj bi izboljšalo doslednost podatkov.
- Več deklarativnih omejitev, dosledna uporaba glavnih ključev ter spoštovanje normalizacije naj bi izboljšalo popolnost podatkov.

Proučevanje povezav med posameznimi razsežnostmi kakovosti podatkov ter posameznimi lastnostmi FIRPM je mogoče delno zaslediti že v literaturi, vendar so avtorji največkrat osredotočeni le na posamezno lastnost FIRPM, medtem ko v knjigi poskušam zgraditi model, ki posamezno razsežnost kakovosti podatkov opiše s funkcijo lastnosti FIRPM. Nekateri avtorji

navajajo vpliv posamezne lastnosti FIRPM na kakovost podatkov, pri čemer ne izpostavijo posamezne razsežnosti kakovosti podatkov, drugi pa tudi povejo, na katero razsežnost kakovosti podatkov posamezna lastnost FIRPM vpliva. Vsem pa je skupno to, da izpostavijo **pozitivno povezavo** med lastnostmi FIRPM in kakovostjo podatkov.

Scannapieca, Missier in Batini (2005) govorijo o deklarativnih omejitvah kot o pomembnih dejavnikih, ki vplivajo na razsežnost »doslednost«. Temu mnenju se pridružujejo tudi Rudra in Yeo (1999) ter Ambler (2006). Ugotavljajo, da je povezava med deklarativnimi omejitvami in doslednostjo pozitivna. Batini in Scannapieca (2006) poudarjata pomembnost tujih ključev za doslednost. Rudra in Yeo (1999) se ukvarjata s proučevanjem normalizacije in ugotovita, da normalizacija pozitivno vpliva na doslednost. O pozitivni povezanosti med doslednostjo in glavnimi ključi pa piše Naumann (2001), ki se osredotoča na primerno izdelavo glavnih ključev v kontekstu izvajanja poizvedb, kar po njegovem mnenju vpliva na doslednost oziroma kakovost podatkov.

Ambler (2003, 2006) predstavlja pozitivno povezanost med deklarativnimi omejitvami in podatkovnimi tipi ter kakovostjo podatkov, modeliranje v knjigi pa je to povezanost natančneje opredelilo, in sicer je nakazalo pozitivno povezanost z razsežnostjo natančnost. V svojih delih priporoča preizkušanje modela zbirke podatkov že pri izdelavi.

Glavni ključi pozitivno vplivajo tudi na popolnost z vidika vrednosti NULL/NOT NULL. O tem sicer ni pisal noben avtor, vendar je bilo to nakazano z modeliranjem. Sta pa Scannapieca in Batini (2006) pisala o vlogi deklarativnih omejitev za popolnost podatkov. Po njunem mnenju je povezava med deklarativnimi omejitvami in popolnostjo pozitivna. Pomen normalizacije glede vrednosti NULL/NOT NULL ter posledično popolnosti pa je poudaril Date (1995), ki pravi, da je z ustrezno normalizacijo mogoče odpraviti vrednosti NULL.

Proučevanje literature in modeliranje nakazujeta **pozitivno** povezanost med lastnostmi FIRPM in kakovostjo podatkov. Knjiga pa ima še globlji namen, in sicer poiskati ustrezno funkcijo. Modeliranje in intuicija narekujeta spoznanje, da je ustrezna funkcija linearna funkcija, z ustreznim modeliranjem se namreč premosorazmerno poveča tudi kakovost podatkov. Zato predpostavljam naslednji linearni model, ki ga bom v nadaljevanju poskušal preizkusiti s statistično analizo:

$D_{21} = k(P_1, P_2) \rightarrow D_{21} = a + b_1 * P_1 + b_2 * P_2,$
$D_{22} = k(P_1, P_2, P_3, P_4, P_5) \rightarrow D_{22} = a + b_1 * P_1 + b_2 * P_2 + b_3 * P_3 + b_4 * P_4 + b_5 * P_5,$

$D_{23} = k(P_2, P_3, P_4) \rightarrow D_{23} = a + b_2 * P_2 + b_3 * P_3 + b_4 * P_4,$

kjer je D_{21} – natančnost, D_{22} – doslednost, D_{23} – popolnost, P_1 – podatkovni tipi, P_2 – deklarativne omejitve, P_3 – normalizacija, P_4 – glavni ključi, P_5 – tuji ključi.

9.4 Preizkušanje modela s statistično analizo

Osnovni namen statistične obdelave podatkov, dobljenih z anketo, je dvojen. Vpliv relacijskega podatkovnega modela na kakovost podatkov, ki je predstavljen z modelom, želim potrditi tudi s statistično analizo. Statistično želim potrditi naslednje trditve:

- Normalizacija vpliva na doslednost in popolnost.
- Podatkovni tipi vplivajo na natančnost in doslednost.
- Deklarativne omejitve vplivajo na natančnost, popolnost in doslednost.
- Tuji ključi vplivajo na doslednost.
- Glavni ključi vplivajo na popolnost in doslednost.
- Število polj v glavnem ključu vpliva na natančnost, doslednost in popolnost.

Iz izpeljanega modela ter proučevane literature sledi:

- **Na natančnost pozitivno vplivajo deklarativne omejitve in podatkovni tipi.**
- **Na popolnost pozitivno vplivajo normalizacija, deklarativne omejitve in glavni ključi.**
- **Na doslednost pozitivno vplivajo deklarativne omejitve, tuji in glavni ključi, normalizacija ter podatkovni tipi.**

Poleg dokazovanja omenjenih trditev bom prikazal tudi nekatere opisne statistike podatkov, ki bodo omogočale izpeljavo ustreznih sklepov glede izbranega raziskovalnega problema.

Iz dosedanjega proučevanja sledi, da je zveza med lastnostmi relacijskega podatkovnega modela in razsežnostmi kakovosti podatkov samo enosmerna, torej lastnosti relacijskega podatkovnega modela lahko vplivajo na razsežnosti kakovosti podatkov, nasprotno pa ne. Dogodek A (podatkovni model) se namreč vedno zgodi pred dogodkom B (kakovost podatkov), zato je za statistično dokazovanje popolnoma dovolj izračun koeficienta korelacije, kar izhaja iz kavzalnega/vzročnega pristopa. Čeprav trditve o vplivu relacijskega podatkovnega modela na kakovost podatkov lahko sprejmem že na podlagi modeliranja s SUBP SQL 2005, jih želim preveriti tudi statistično. V primeru tudi statistične potrditve izvedenih trditev bo sprejetje glavne trditve lažje in bolj utemeljeno.

9.4.1 Vzorec združb

V preglednici 17 je opisan vzorec vrnjenih in pravilno izpolnjenih anketnih vprašalnikov glede na velikost združbe po številu zaposlenih, iz katerega je mogoče sklepati, da vzorec ni

reprezentativen za slovenske združbe, saj je v vzorcu največ velikih in srednje velikih združb, kar je ravno nasprotno od dejanske sestave populacije.

PREGLEDNICA 17: IZPOLNJENI ANKETNI VPRAŠALNIKI GLEDE NA VELIKOST ZDRUŽB IN PRIMERJAVA S POPULACIJO

Velikost združbe po številu zaposlenih	**Število združb v anketnem vprašalniku**	**Delež v anketnem vprašalniku (v %)**	**Populacija združb v letu 2006**	**Delež v populaciji v letu 2006 (v %)**
Mikro (0–9)	8	10,8	93.430	92,9
Majhna (10–49)	17	23,0	5601	5,6
Srednje velika (50–249)	22	29,7	1255	1,2
Velika (250+)	27	36,5	283	0,3
Skupaj	74	100	100.569	100

Začetno predvidevanje, da bom največ vrnjenih izpolnjenih anketnih vprašalnikov dobil od združb, ki imajo lastno službo informatike, se je uresničilo, saj je med anketiranimi združbami kar 87,8 % takšnih, ki imajo lastno službo za informatiko. Analiza podatkov pokaže tudi povezanost med velikostjo združbe in lastno službo za informatiko, saj je vrednost preverjanja hi kvadrat 36,833 pri točni stopnji značilnosti 0,000.

Koeficient korelacije med številom zaposlenih v združbi in številom zaposlenih v službi za informatiko je statistično značilen in znaša 0,234 pri točni stopnji značilnosti 0,044. Iz povedanega sledi, da je večje število srednje velikih in večjih združb med anketiranimi združbami logična posledica obravnavanega raziskovalnega problema. Vendar je obravnavano raziskovalno vprašanje neodvisno od velikosti združb, kar lahko potrdim tudi s kontingenčno preglednico. V preglednici 18 so upoštevani le jasno opredeljeni odgovori na vprašanje o težavah s kakovostjo podatkov, odgovor »Ne vem« pa je izvzet.

PREGLEDNICA 18: KONTINGENČNA PREGLEDNICA – VELIKOST ZDRUŽB TER KAKOVOST PODATKOV

Hi kvadrat (vrednost 4,172; točna stopnja značilnosti 0,243)		Težave zaradi (ne)kakovosti podatkov	
		Da	Ne
Velikost združbe po številu zaposlenih	Mikro	2	6
	Majhna	10	6
	Srednje velika	13	7
	Velika	13	12

Preverjanje hi kvadrat za obravnavani spremenljivki velikost združbe in težave zaradi (ne)kakovosti podatkov nam pokaže, da spremenljivki nista povezani, zato lahko sprejmem

trditev, da težava s kakovostjo podatkov lahko doleti vsako združbo, neodvisno od njene velikosti po številu zaposlenih.

Da bi dodatno ovrgel sum o odvisnosti raziskovalnega problema od velikosti združb, sem naredil še kontingenčne preglednice za spremenljivki »velikost združbe« in »natančnost podatkov«, spremenljivki »velikost združbe« in »popolnost podatkov«, spremenljivki »velikost združbe« in »doslednost podatkov« ter spremenljivki »velikost združbe« in »zaupanje v podatke«.

PREGLEDNICA 19: POVEZANOST MED VELIKOSTJO ZDRUŽB IN RAZSEŽNOSTMI KAKOVOSTI PODATKOV

	Hi kvadrat	
Spremenljivki	Vrednost	Točna stopnja značilnosti
Velikost združbe in natančnost podatkov	8,230	0,511
Velikost združbe in doslednost podatkov	14,135	0,292
Velikost združbe in popolnost podatkov	13,460	0,337
Velikost združbe in zaupanje v podatke	16,811	0,052

V preglednici 19 sta prikazani vrednosti preverjanja hi kvadrat za omenjene spremenljivke. Iz preglednice je jasno razvidno, da nobena razsežnost kakovosti podatkov ni povezana z velikostjo združbe.

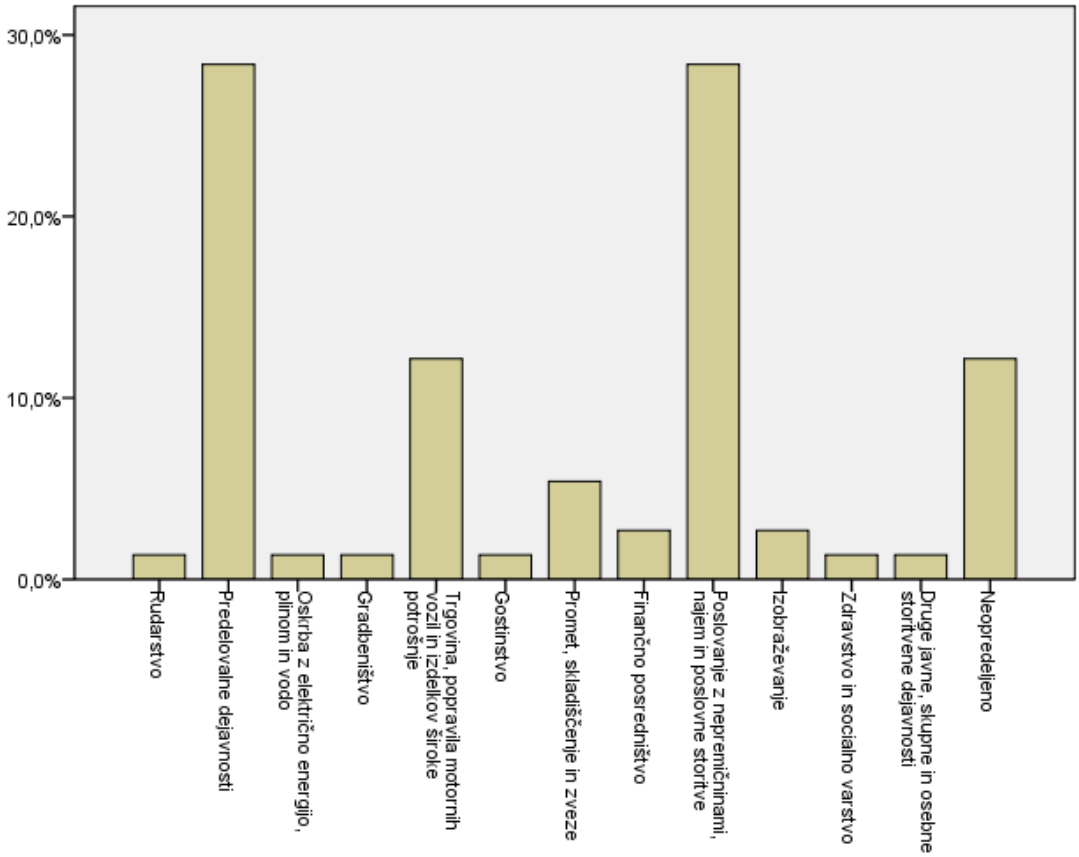
Prav tako naredim kontingenčne preglednice za spremenljivki »velikost združbe« in »tuji ključi«, spremenljivki »velikost združbe« in »deklarativne omejitve«, spremenljivki »velikost združbe« in »normalizacija« ter spremenljivki »velikost združbe« in »podatkovni tipi« (preglednica 20).

Med velikostjo združb in proučevanimi lastnostmi FIRPM statistično ne morem potrditi povezanosti. S proučevanjem sem torej ovrgel povezanost med velikostjo združbe in težavami zaradi nekakovostnih podatkov, med razsežnostmi natančnost, popolnost, doslednost in zaupanje ter med velikostjo združbe in proučevanimi lastnostmi FIRPM. Problem kakovosti podatkov je torej neodvisen od velikosti združb, kar je dokazal že Lee s soavtorji (2002) in čemur je pritrdila tudi moja raziskava, zato v nadaljevanju velikost združbe ni več pomemben dejavnik.

PREGLEDNICA 20: POVEZANOST MED VELIKOSTJO ZDRUŽB IN FIRPM

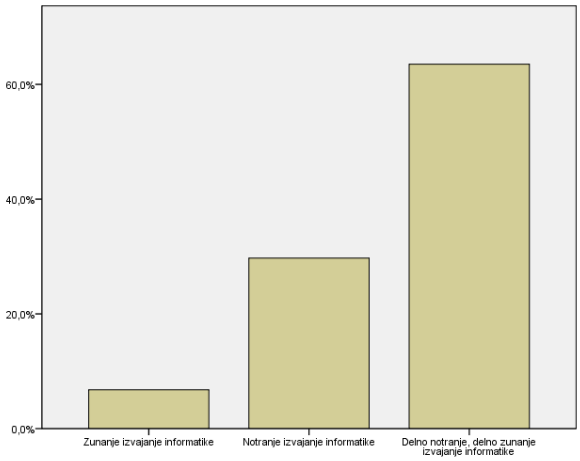
	Hi kvadrat	
Spremenljivki	Vrednost	Točna stopnja značilnosti
Velikost združbe in tuji ključi	11,069	0,523
Velikost združbe in deklarativne omejitve	12,432	0,412
Velikost združbe in normalizacija	10,235	0,332
Velikost združbe in podatkovni tipi	11,062	0,086

SLIKA 28: ANKETIRANE ZDRUŽBE PO PANOGAH DEJAVNOSTI



Na sliki 28 je prikazana razporeditev anketiranih združb po panogah dejavnosti.

SLIKA 29: ANKETIRANE ZDRUŽBE PO UREJENOSTI INFORMATIKE



Na podlagi opravljenih izračunov lahko trdim, da za dokazovanje vpliva FIRPM na kakovost podatkov ne potrebujem reprezentativnega vzorca združb z vidika velikosti združb, temveč le

zadostno število statističnih enot, kar pa je v mojem primeru (število enot 74) izpolnjeno. V nadaljevanju bom za združbe, ki so vrnile pravilno izpolnjen anketni vprašalnik, uporabljal izraz anketirane združbe. Anketirane združbe so odgovarjale še na dve splošni vprašanji, od katerih je bilo eno neobvezno, drugo pa obvezno. Neobvezno vprašanje se je nanašalo na panogo dejavnosti, navesti je bilo treba samo prvo črko v standardni klasifikaciji dejavnosti (glej SURS, spletna stran za klasifikacije), v kateri združba deluje, obvezno vprašanje pa se je nanašalo na urejenost informatike v združbi.

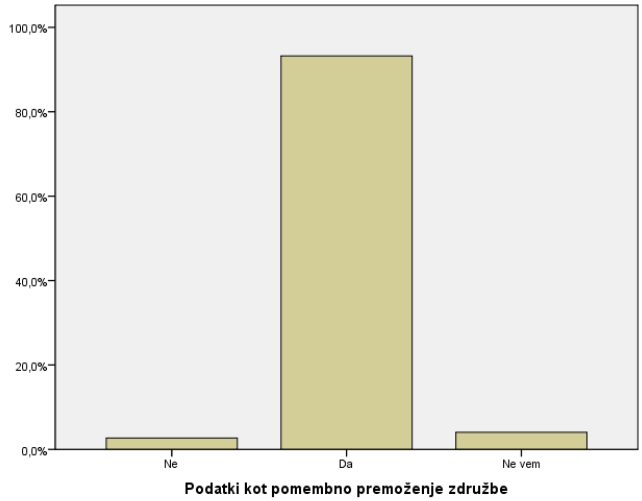
Med anketiranimi združbami prevladujeta panogi »Predelovalne dejavnosti« ter »Poslovanje z nepremičninami, najem in poslovne storitve« z 28,4 %, sledita pa ji panoga »Neopredeljeno« ter panoga »Trgovina, popravila motornih vozil in izdelkov široke potrošnje« z 12,2 %. Vprašanje o urejenosti informatike v združbi se je nanašalo na urejenost informatike v razmerju do zunanjega izvajanja informatike. Največ anketiranih združb, kar 64 %, je imelo združek lastne informatike ter zunanjega izvajanja informatike, 30 % združb je imelo lastno informatiko, ostali delež združb pa se je zanesel samo na zunanje izvajanje informatike (slika 29).

S tem sem končal opisovati vzorec, v nadaljevanju pa sledi obdelava drugega dela anketnega vprašalnika.

9.4.2 Osnovne opisne statistike in primerjava z mednarodnima raziskavama

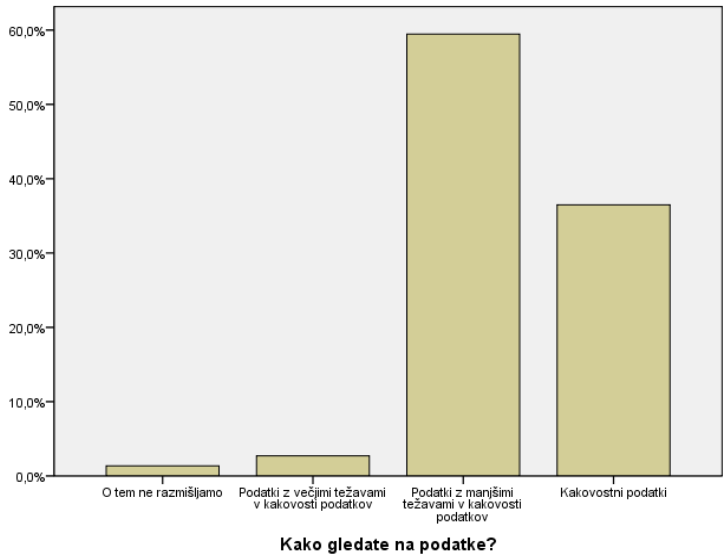
V drugem delu anketnega vprašalnika sem se osredotočil na raziskovanje težav s kakovostjo podatkov v združbah. Zanimalo me je, ali združbe razpravljajo o kakovosti podatkov, kako združbe gledajo na podatke in kdo je pobudnik razprav o kakovosti podatkov v združbah.

SLIKA 30: PODATKI KOT PREMOŽENJE ZDRUŽBE



Iz upodobitve na sliki 30 je jasno razvidno, da večina anketiranih združb (93,2 %) gleda na svoje podatke kot na pomembno premoženje združbe. Če omenjeni delež združb primerjam z Amblerjevo (2006) raziskavo, vidim, da je delež združb, ki na podatke gledajo kot na premoženje, v Sloveniji nekoliko manjši kot v ZDA, kjer je delež združb 96-odstoten, vendar med deležema ni velikih odstopanj.

SLIKA 31: ZAZNAVANJE KAKOVOSTI PODATKOV V ZDRUŽBAH



Čeprav večina združb pozitivno gleda na svoje podatke, pa ima več kot 60 % anketiranih slovenskih združb večje ali manjše težave s kakovostjo podatkov. 62,1 % anketiranih združb na podatke gleda kot na manjše ali večje težave v kakovosti podatkov, kot kakovostne podatke pa svoje podatke ocenjuje 36,5 % anketiranih združb (slika 31).

V ZDA po raziskavi TDWI (Russom 2006) kot kakovostne podatke ter podatke, ustrezne za nameravano uporabo, ocenjuje 48 % združb, vendar le 10 % združb nanje gleda kot na popolnoma kakovostne podatke. V slovenskih anketiranih združbah je ta delež večji in znaša 35 %. Razlika je najbrž posledica obdobja, v katerem sta se izvajali anketi (v ZDA leta 2005), saj se je v tem obdobju sunkovito povečala vloga poslovnega obveščanja, ki zahteva kakovostne podatke.

Nekatere anketirane združbe so navedle konkretne vzroke za nekakovostne podatke. Konkretne vzroke je navedlo 21 združb, njihove navedbe pa sem uvrstil v eno od kategorij raziskave TDWI iz leta 2005 (Russom 2006). Primerjava je prikazana v preglednici 21. Zavedam se, da je za verodostojnešo primerjavo premalo odgovorov, vendar je iz dobljenih odgovorov že mogoče čutiti podobnost z raziskavo TDWI. Na podlagi navedenega lahko sklepam, da so slovenske anketirane združbe primerljive z anketiranimi združbami v ZDA.

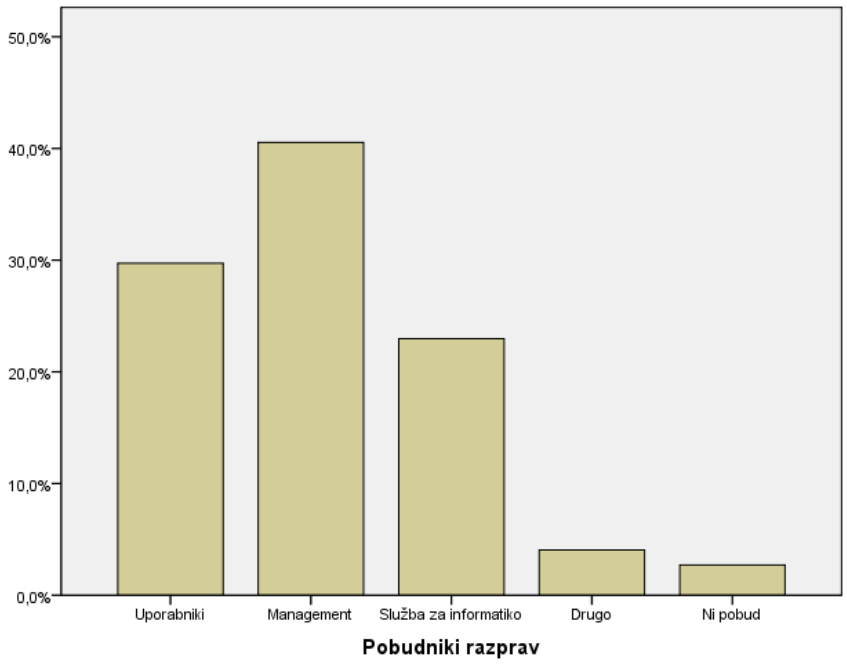
PREGLEDNICA 21: VZROKI ZA NEKAKOVOST PODATKOV

Najpogostejši vzroki za nekakovost podatkov v združbah v ZDA (od najbolj pogostega do najmanj pogostega)	Najpogostejši vzroki za nekakovost podatkov v združbah v Sloveniji
Nepopolne opredelitve posameznih entitet	23,8 %
Napačen vnos podatkov zaposlenih	4,8 %
Različni prehodi in obdelave podatkov	23,8 %
Različna pričakovanja uporabnikov	14,3 %
Zunanji podatki	4,8 %
Napačen vnos podatkov strank	4,8 %
Sistemske napake	9,6 %
Spremembe vhodnih postopkov	4,8 %
Drugo	4,8 %

Vir: Lastni preračuni; Russom, Liability and leverage: a case for data quality, 2006.

Na področju kakovosti podatkov je veliko prispeval Olson (2003), ki je med drugim uvedel v knjigi že pojasnjena pojma »od spodaj navzgor« in »od zgoraj navzdol«. Prvi pojem opisuje razmere, ko pobude za razpravo o kakovosti podatkov prihajajo od uporabnikov in službe za informatiko, drugi pojem pa te pobude pripisuje managementu.

SLIKA 32: POBUDNIKI RAZPRAV IN UKREPOV GLEDE KAKOVOSTI PODATKOV



Med anketiranimi združbami večina pobud za razpravo o kakovosti podatkov ter sprejemanje različnih ukrepov prihaja »od spodaj navzgor», torej od uporabnikov (29,7 %) in službe za

informatiko (23 %), medtem ko je v 40,5 % anketiranih združbah pobudnik razprav management (slika 32). Smiselno bi bilo spremljati pobudnike razprav v združbah v prihodnosti in najbrž se bo delež združb, kjer pobude o kakovosti podatkov prihajajo od managementa, še povečal. To domnevo lahko postavim glede na povečevanje uporabe poslovnega obveščanja in odvisnosti pomembnosti kakovosti podatkov od zapletenosti analiz (glej slika 20). Ker so zapletene analize po navadi del dolgotrajnih projektov, ki so tudi finančno obremenjujoči za posamezno združbo, mora management storiti vse potrebno, da takšni projekti uspejo, torej mora zagotoviti tudi kakovostne podatke. Tega ne more storiti neposredno, zato je njegova naloga zagotoviti razmere, ki bodo omogočale kakovostne podatke. Podobne so ugotovitve že večkrat omenjene raziskave TDWI, kar pomeni, da se slovenske anketirane združbe glede kakovosti podatkov ne razlikujejo pomembno od združb v informacijsko bolj razvitih ZDA.

9.4.3 Obdelava odgovorov o kakovosti podatkov

V preglednici 22 sta prikazana povprečje in standardni odklon vprašanj o kakovosti podatkov. Med vsemi razsežnostmi kakovosti podatkov je razsežnost »popolnost« dobila najslabšo oceno (3,92), najboljšo oceno (4,31) pa je dobila razsežnost »zaupanje«.

PREGLEDNICA 22: OPISNA STATISTIKA ODGOVOROV O KAKOVOSTI PODATKOV

Spremenljivka	Povprečje	Standardni odklon
Natančnost podatkov	4,22	0,647
Doslednost podatkov	4,11	0,885
Popolnost	3,92	0,903
Zaupanje v podatke	4,31	0,720
Pomembnost poslovnega informacijskega sistema za poslovanje združbe	4,41	0,701
Ali pri odločitvi za nakup PIS kakovost zbirke podatkov igra pomembno vlogo?	4,07	0,896
Preverjanje podatkov v drugih podatkovnih virih	3,31	1,006
Posledice odločitev na podlagi podatkov iz informacijskega sistema so v skladu s pričakovanji.	4,15	0,612
Kako pogosto dostopate do podatkov v informacijskem sistemu tudi z drugimi orodji, npr. Accessom, Excelom?	2,92	1,269
Kako pogosto popravljate podatke neposredno v zbirki podatkov?	4,07	0,865

Legenda: v poševnem tisku so navedene merjene razsežnosti kakovosti podatkov

Razsežnost »zaupanje« je dobila visoko oceno, pa vendar preseneča povprečna vrednost preverjanja v drugih virih podatkov. Čeprav anketirane združbe močno zaupajo v podatke, še vedno preverjajo podatke v drugih podatkovnih virih. Spremenljivki tudi statistično nista povezani (vrednost koeficienta korelacije je 0,130, točna stopnja značilnosti pa je 0,271). Očitno

je, da anketirane združbe mislijo, da zaupajo podatkom, v resnici pa opravijo še kar nekaj preverjanj, da se prepričajo, da je podatek v PIS vreden zaupanja.

PREGLEDNICA 23: IZRAČUN KAKOVOSTI PODATKOV V ANKETIRANIH SLOVENSKIH ZDRUŽBAH

$K_2 = [D_{21} * D_{22} * D_{23} * D_{24}, \min(D_{21}, D_{22}, D_{23}, D_{24})]$		
Razsežnost kakovosti podatkov	**Vrednost razsežnosti**	**Normalizirana vrednost razsežnosti**
D_{21} – natančnost	4,22	0,84
D_{22} – doslednost	4,11	0,82
D_{23} – popolnost	3,92	0,78
D_{24} – zaupanje	4,31	0,86
$Q = K_2$	[0,46–0,78]	

Ker imam znane vrednosti posameznih razsežnosti, lahko za anketirane združbe izračunam kakovost podatkov z enačbo na strani 109. V tem primeru je $Q = K_2$, ker proučujem kakovost podatkov samo v koraku shranjevanja podatkov, torej v PIS, za namen uporabe (U_j) pa predpostavljam vrednost 0. V preglednici 23 so prikazane merjene razsežnosti kakovosti podatkov v anketiranih združbah. Poleg razsežnosti kakovosti podatkov, ki sem jih modeliral, sem v preglednico 23 dodal še razsežnost zaupanje.

Izračun pokaže, da je kakovost podatkov anketiranih slovenskih združb po najbolj pesimistični oceni 0,46, po optimistični oceni pa 0,78.

PREGLEDNICA 24: KOEFICIENT KORELACIJE MED POSLEDICAMI ODLOČITEV IN RAZSEŽNOSTMI KAKOVOSTI PODATKOV

Spremenljivka	**Koeficient korelacije**	**Točna stopnja značilnosti (F-test)**
Posledice odločitev na podlagi podatkov iz informacijskega sistema so v skladu s pričakovanji.	0,644	0,000
Natančnost		
Posledice odločitev na podlagi podatkov iz informacijskega sistema so v skladu s pričakovanji.	0,552	0,000
Doslednost		
Posledice odločitev na podlagi podatkov iz informacijskega sistema so v skladu s pričakovanji.	0,518	0,000
Popolnost		

Kako pomembni so kakovostni podatki za odločanje, pokaže povezanost med razsežnostmi natančnost, doslednost in popolnost podatkov ter sprejemanjem odločitev. Bolj natančni, dosledni in popolni so podatki, lažje je predvideti posledice naših odločitev. Za dokazovanje povezanosti med navedenimi spremenljivkami sem izračunal koeficient korelacije (preglednica 24). Vse povezave so pozitivne in dokaj močne. Čeprav koeficient korelacije v osnovi pove samo

jakost in smer povezave dveh spremenljivk, lahko na podlagi teorije in modela sklepam na smer odvisnosti. Logično je, da so podatki osnova za odločitve, torej se časovno vedno zgodijo pred sprejemanjem odločitev, zato lahko sklepam, da kakovost podatkov vpliva na odločanje, nasprotno pa ne.

Smiselno je preveriti, kako je razsežnost zaupanje povezana z razsežnostmi natančnost, popolnost in doslednost.

Ker Pradhan (2005) ter Wang in Strongova (1996) govorijo o zaupanju v podatke kot o zaupanju v natančnost, resničnost, doslednost, izračunam še povezanost kot linearno regresijo, kjer je »zaupanje« odvisna spremenljivka, »natančnost« ter »doslednost« pa neodvisni spremenljivki (preglednica 25). Regresijska analiza je pokazala, da bo zaupanje v podatke večje, če so podatki natančni in dosledni, kar je popolnoma logično. Izvedena regresijska analiza mi je dala le površni pregled vpliva na zaupanje. Ker je zaupanje psihološki dejavnik, je preveč zapleten, da bi ga mogel opisati zgolj z dvema neodvisnima dejavnikoma, a tudi omenjena površna ocena ni tako zgrešena. Popravljen R^2 nam v navedenem primeru pove, da je 35 % variance zaupanja pojasnjenih z linearnim vplivom natančnosti in doslednosti. Regresijska analiza torej pove, da se bo zaupanje uporabnikov v podatke povečalo, če se bodo združbe trudile, da bodo podatki v njihovih PIS bolj natančni in dosledni.

PREGLEDNICA 25: REGRESIJSKA ANALIZA ZA ODVISNO SPREMENLJIVKO »ZAUPANJE V PODATKE«

Popravljen R^2 = 0,351 F = 20,741 (točna stopnja značilnosti 0,000)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	1,598	0,453	3,525	0,001
Doslednost podatkov	0,296	0,094	3,134	0,003
Natančnost podatkov	0,355	0,129	2,755	0,007
Odvisna spremenljivka: Zaupanje v podatke				

PREGLEDNICA 26: KONTINGENČNA PREGLEDNICA – VELIKOST ZDRUŽB TER POPRAVLJANJE PODATKOV

Spremenljivki		Kako pogosto popravljate podatke neposredno v zbirki podatkov?				
		2	*3*	*4*	*5*	Skupaj
Velikost združbe po številu zaposlenih	*Mikro*	2	2	2	2	8
	Majhna	0	2	8	7	17
	Srednje velika	1	4	8	9	22
	Velika	1	5	13	8	27
	Skupaj	4	13	31	26	74

Presenetil me je rezultat vprašanja o popravljanju podatkov neposredno v zbirki podatkov, saj sem bil prepričan, da bo slabši. Povprečje 4,07 pomeni, da anketirane združbe redko popravljajo podatke neposredno v zbirki podatkov. Verjetno so omenjene združbe spoznale, da je to početje tvegano, omenjeni rezultat pa je najbrž tudi posledica, da je med anketiranimi združbami največ velikih združb, v katerih verjetno prevladujejo svetovno najbolj priznani PIS (na primer SAP, Navision) in kjer neposredni, nepooblaščeni dostop do podatkov pomeni izgubo jamstva. Podrobnejši pregled (preglednica 26) pokaže, da spremenljivki »velikost združbe« ter »popravljanje podatkov neposredno v zbirki podatkov« nista povezani. Preverjanje hi kvadrat namreč ne pokaže statistične značilnosti, saj je vrednost točne stopnje značilnosti 0,411.

Podrobneje je treba proučiti tudi povezanost med pomembnostjo PIS za poslovanje združbe ter vlogo kakovosti zbirke podatkov pri nakupu PIS (preglednica 27).

Pričakoval bi, da sta spremenljivki pozitivno povezani, torej bi z naraščanjem vloge PIS za poslovanje združbe morala naraščati tudi vloga zbirke podatkov, a izračun koeficienta korelacije pokaže, da sta spremenljivki šibko povezani, vendar povezava statistično ni značilna. Čeprav sta obe spremenljivki dobili visoko povprečno vrednost, povezanosti med njima ne morem potrditi.

PREGLEDNICA 27: KOEFICIENT KORELACIJE MED POMEMBNOSTJO PIS IN POMEMBNOSTJO ZBIRKE PODATKOV

Spremenljivka	Koeficient korelacije	Točna stopnja značilnosti (F-test)
Ali pri odločitvi za nakup informacijskega sistema kakovost zbirke podatkov igra pomembno vlogo?	0,130	0,269
Pomembnost poslovnega informacijskega sistema za poslovanje združbe		

Na žalost omenjeni izračun potrdi domnevo, da anketirane združbe pri nakupu PIS neustrezno obravnavajo zbirko podatkov. Pri vseh odločitvah je treba upoštevati vidik nameravane uporabe, torej če je PIS ključen za poslovanje združbe, bi morala združba pri nakupu ustrezno proučiti tudi kakovost podatkovnega modela, ki je del PIS.

Ena od lastnih izkušenj, ki so zapisane tudi v uvodnem poglavju knjige, je bila, da združbe pogosto dostopajo do podatkov v PIS tudi z drugimi orodji. Obdelava anket je pokazala, da se nisem motil, saj je povprečna vrednosti te spremenljivke 2,92. Omenjeno povprečje napeljuje k sklepu, da anketirane združbe očitno uporabljajo PIS, ki njihovih potreb ne zadovoljuje popolnoma, zato so prisiljene uporabiti tudi druga orodja. Za podkrepitev takšnega sklepa bi sicer potreboval še odgovor na dodatno vprašanje, in sicer ali PIS omogoča izvoze v druga programska orodja in tako dopolni svoje funkcije ali tega ne omogoča in združbe uporabljajo druga programska orodja po lastni presoji. Kot projektant in programer vem, da je v PIS skoraj

nemogoče vgraditi toliko različnih analiz, izpisov, kolikor je želja končnih uporabnikov. Lažje je narediti izvoz podatkov v Excel, na primer, in pustiti končnemu uporabniku, da v Excelu nadaljuje z različnimi analizami podatkov. Druga možnost, ki je čedalje bolj priljubljena, je uporaba sicer omejeno zmogljivih vrtilnih preglednic v SSOP, kar močno zmanjša potrebo po različnih izpisih.

9.4.4 Obdelava odgovorov o kakovosti zbirke podatkov

V preglednici 30 sta prikazana povprečje ter standardni odklon vprašanj o kakovosti zbirke podatkov. Na podlagi izračuna lahko trdim, da ima povprečni podatkovni model PIS v anketiranih združbah naslednje lastnosti: skoraj vse preglednice imajo glavne ključe, glavni ključi so sestavljeni iz nekaj polj, delno denormaliziran podatkovni model, tuji ključi sicer so navzoči, vendar manj, kot bi pričakoval. Presenetljivo pogosta je uporaba deklarativnih omejitev, medtem ko ustrezna uporaba podatkovnih tipov ne preseneča, saj je njihova ustreznost pomembna za kakovosten podatkovni model. Čeprav je uporaba glavnih ključev izredno velika, sem pričakoval, da je ta še večja. Očitno ima povprečna anketirana združba eno ali nekaj preglednic, ki jih uporablja kot preglednice za uvoz velikih količin podatkov, kjer bi glavni ključi upočasnili proces uvoza. Opredelitev glavnega ključa namreč povzroči ustvarjenje indeksa, ki se mora pri vsakem uvoženem zapisu posodobiti, kar vpliva na hitrost uvoza. Še enkrat pa je treba poudariti, da je tudi uporaba tujih ključev manjša od pričakovane, vendar po drugi strani dobljena vrednost te spremenljivke potrjuje način prehoda na novejše platforme. Pred razvojem sodobnega SUBP so razvijalci PIS uporabljali različne zbirke podatkov. S pojavom sodobnega SUBP so omenjeni razvijalci želeli na enostaven način prestaviti svoje podatkovne modele na novejše platform in razvijalci SUBP so jim ugodili.

PREGLEDNICA 28: OPISNA STATISTIKA ODGOVOROV O KAKOVOSTI ZBIRKE PODATKOV

Spremenljivka	Povprečje	Standarndni odklon
Normalizacija	3,86	0,751
Tuji ključi	3,19	1,449
Glavni ključi	4,50	0,806
Iz koliko polj je v večini primerov sestavljen glavni ključ?	3,54	1,149
Deklarativne omejitve	3,57	0,994
Podatkovni tipi	4,47	0,624
Za shranjevanje številčnih podatkov so uporabljena splošna znakovna polja.	4	1

V razvoj sodobnejših, robustnejših podatkovnih modelov je bilo vloženega premalo časa in truda, zato so podatkovni modeli mnogih razvijalcev SUBP s seboj prinesli vse pomanjkljivosti iz preteklosti. Razvijalce PIS očitno čaka še ogromno dela pri posodabljanju in izboljšavi podatkovnih modelov. Omenjen izračun je pokazal tudi na dejstvo, da razvijalci PIS za številčne podatke, sicer redko, pa vendar uporabljajo tudi znakovne podatkovne tipe. To nakazuje na dejavnosti, da bi razvijalci PIS lahko hitro ugodili uporabnikom PIS glede njihovih zahtev po novih funkcijah, kar jih sili v »zlorabo« polj ali dodajanje polj v podatkovni model »na zalogo«. Zadeva je dokaj preprosta. Uporabniki niso pripravljeni dolgo čakati na izpolnitev določene želje oziroma zahteve, zato so razvijalci prisiljeni iskati hitre rešitve. Nadgradnja podatkovnega modela ni preprosta naloga, kar lahko potrdim tudi iz lastnih izkušenj, zato je za določen postopek, funkcijo laže izkoristiti kakšno splošno polje. V dveh združbah, v katerih sem delal kot projektant in programer in ki sta bili razvijalki PIS je razvoj PIS potekal po naslednjem postopku. V enem letu se je PIS po navadi dvakrat redno nadgradil in nadgradnji sta vsebovali obsežnejše nove funkcije. Vendar je bilo med načrtovanimi nadgradnjami veliko želja in pritiski po izrednih nadgradnjah so bili izjemno veliki. Pritiskom se je mnogokrat tudi popustilo, še posebno če so bili to pritiski pomembnih končnih uporabnikov. Zato je bila včasih končnemu uporabniku poslana samo izvršilna datoteka (exe) ali pa določena knjižnica (dll), ki je vsebovala novo funkcijo, podatkovni model pa se ni spremenil, čeprav bi ga bilo treba za določene funkcije spremeniti, vendar se je s spremembo počakalo na redno nadgradnjo. V tem primeru ni bilo drugega izhoda kot »zloraba« že dodanega polja in nameravana kratkotrajna »zloraba« je včasih postala kar trajna »zloraba«. S sodobnim SUBP je takšne zaplete oziroma zahteve mogoče elegantneje rešiti. V SQL 2005, na primer, obstaja možnost uporabe zapletenega podatkovnega tipa XML, v katerega lahko dodamo tako rekoč neomejeno število novih polj, za katere lahko opredelimo tudi shemo, ki nadzira vnose v omenjena polja, shranjena v podatkovnem tipu XML.

Končna ocena kakovosti FIRPM v anketiranih združbah je sicer zadovoljiva, vendar obstaja še kar nekaj prostora za izboljšave, za kar bi si morali razvijalci PIS tudi prizadevati.

Orr (1998) govori, da se kakovost podatkov v PIS poslabša s starostjo PIS. V svojem delu ni izpostavil, katera razsežnost kakovosti podatkov se poslabša s starostjo PIS, zato sem v preglednici 31 izračunal koeficiente korelacije med starostjo PIS in razsežnostmi natančnost, popolnost in doslednost. Vendar pa je treba njegovo trditev proučiti tudi s stališča časa. V 90. letih prejšnjega stoletja so si namreč relacijski podatkovni modeli šele utirali pot na trg, kar posledično pomeni, da je Orrovo proučevanje zajelo tudi PIS, ki so za podatkovni model uporabljali proste preglednice oziroma datotečne sisteme. V opravljeni anketi pa so zajeti novejši sistemi (povprečna starost PIS je 6,86 leta), ki uporabljajo relacijski podatkovni model. Poudariti je treba še eno veliko razliko med časom Orrove raziskave in sodobnimi sistemi. Danes je običajno, da združbe v PIS hranijo le podatke za tekoče leto, starejše podatke pa prenesejo v

podatkovna skladišča, kar je bilo včasih zaradi visokih cen programske in strojne opreme neizvedljivo in so PIS hranili podatke za daljše obdobje.[15]

PREGLEDNICA 29: KOEFICIENTI KORELACIJE MED RAZSEŽNOSTMI KAKOVOSTI PODATKOV IN STAROSTJO PIS

Spremenljivke	Natančnost	Doslednost	Popolnost
Starost PIS	0,034 (točna stopnja značilnosti = 0,775)	–0,107 (točna stopnja značilnosti = 0,336)	0,173 (točna stopnja značilnosti = 0,140)

To pomeni, da se določeni podatki niso več oziroma so se redko uporabljali, kar posledično pomeni, da se je kakovost podatkov v PIS zmanjšala. Iz preglednice 29 je razvidno, da koeficient korelacije nakazuje negativno povezanost med doslednostjo in starostjo PIS, vendar povezava ni statistično značilna.

Vendar se nisem zadovoljil samo z izračunom koeficienta korelacije, temveč sem poskusil tudi s SOZP, in sicer z algoritmom *Clustering*. Algoritem je predlagal sedem gruč. V preglednici 30 je kot primer prikazana gruča 5. Zavedati se je treba omejitev opravljenega izračuna, saj je za ustrezno sprejemanje sklepov premalo statističnih enot. Izračun kljub manjšemu številu statističnih enot nakazuje, da najbrž obstaja povezanost med starostjo PIS in kakovostjo podatkov.

PREGLEDNICA 30: ZNAČILNOSTI GRUČE 5

Spremenljivka	Vrednost	Verjetnost
Težave s kakovostjo podatkov	1	47,287 %
Težave s kakovostjo podatkov	0	44,994 %
Starost PIS	10–20,8	40,870 %
Starost PIS	6,9–10	25,285 %
Starost PIS	3,8–6,9	19,249 %
Starost PIS	1–3,8	8,991 %

9.4.5 Obdelava izvedenih trditev

Obravnaval sem vse dele anketnega vprašalnika, zdaj pa je čas, da poskusim potrditi glavno trditev s pomočjo proučevanja izvedenih trditev.

[15] Oracle 8, ki je omogočal zvezdnato shemo, je bil predstavljen leta 1997, njegov tekmec MS SQL Server 7.0 pa leta 1998.

Za potrditev izvedenih trditev je potreben dokaz, da določena lastnost FIRPM vpliva vsaj na eno od razsežnosti kakovosti podatkov.

PREGLEDNICA 31: KOEFICIENTI KORELACIJE MED RAZSEŽNOSTMI KAKOVOSTI PODATKOV IN FIRPM

Spremenljivke	Natančnost	Doslednost	Popolnost
Normalizacija		0,417 (točna stopnja značilnosti =0,000)*	0,425 (točna stopnja značilnosti = 0,000)*
Podatkovni tipi	0,286 (točna stopnja značilnosti = 0,013)*	0,402 (točna stopnja značilnosti = 0,000)*	
Deklarativne omejitve	0,275 (točna stopnja značilnosti = 0,018)*	0,365 (točna stopnja značilnosti = 0,001)*	0,372 (točna stopnja značilnosti = 0,001)*
Glavni ključi		0,146 (točna stopnja značilnosti = 0,213)	0,241 (točna stopnja značilnosti = 0,039)*
Število polj v glavnem ključu		0,020 (točna stopnja značilnosti = 0,869)	0,171 (točna stopnja značilnosti = 0,151)
Tuji ključi		0,208 (točna stopnja značilnosti = 0,075)	

Legenda: * – statistično značilne povezave.

Najprej bom za vse možnosti, ki izhajajo iz modeliranja, izračunal koeficient korelacije. Že večkrat sem poudaril, da je zaradi sestave raziskovalnega problema koeficient korelacije povsem dovolj za potrditev ali zavrnitev teze. V preglednici so izračunani koeficienti korelacije med spremenljivkami, kot izhajajo že iz modeliranja (preglednica 31).

Čeprav sem uspel zbrati le 74 statističnih enot, sem v nadaljevanju opravil tudi analizo z algoritmom *Naive Bayes*. Algoritem izračuna verjetnosti med vhodnimi in izhodnimi spremenljivkami, pri čemer predpostavlja, da so vse spremenljivke med seboj neodvisne. Prav zaradi te predpostavke je algoritem dobil naziv *Naive*, kar pomeni naiven. Njegova prednost je, da za učenje ne potrebuje velikega števila statističnih enot, vendar je 74 kljub temu na spodnji meji. Zato sem se odločil, da bom analizo uporabil samo za potrditev domneve, da podatkovni model (dogodek A) enosmerno vpliva na kakovost podatkov (dogodek B), saj nam izračun pokaže tudi smer vplivanja. Analizo sem opravil s SUBP MS SQL 2005. Algoritem Naive Bayes je del SOZP v MS SQL.

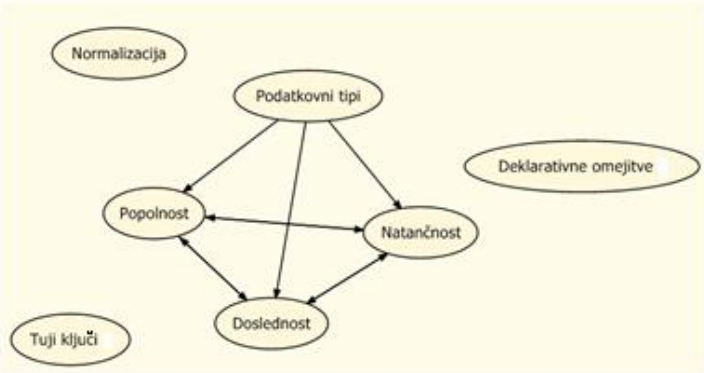
Za določitev smeri odvisnosti sem predpostavil, da nobena razsežnost kakovosti podatkov ne sme vplivati na katero koli lastnost FIRPM, medtem ko mora vsaj ena lastnost FIRPM vplivati na katero koli razsežnost kakovosti podatkov. V tem primeru je smer odvisnosti FIRPM -> kakovost podatkov.

PREGLEDNICA 32: SPREMENLJIVKE V ALGORITMU NAIVE BAYES

Vhodne (neodvisne) spremenljivke	Izhodne (odvisne) spremenljivke
Natančnost	Natančnost
Doslednost	Doslednost
Popolnost	Popolnost
Normalizacija	Normalizacija
Podatkovni tipi	Podatkovni tipi
Deklarativne omejitve	Deklarativne omejitve
Tuji ključi	Tuji ključi

Algoritem ima naslednjo enačbo $p(C|F_1,\ldots,F_n) = \frac{p(C)*p(F_1,\ldots F_n|C)}{p(F_1,\ldots F_n)}$ in najlažje ga je razložiti na zelo preprostem primeru. Verjetnost, na primer, da je oseba, stara 20 let, študent, je 60-odstotna. Verjetnost, da je oseba, ki nosi zapiske, študent, je 85-odstotna. Kolikšna je verjetnost, da je neka oseba, ki je stara 20 let in nosi zapiske, študent? P = 0,6 * 0,85 / (0,6 * 0,85 + (1 – 0,6) * (1 – 0,85)) = 0,895 = 89,5 %. V algoritem sem vnesel spremenljivke, prikazane v preglednici 32, in iste spremenljivke sem opredelil tudi kot izhodne (angl. *predictable*) spremenljivke. Algoritem *Naive Bayes* pokaže (najmočnejše) odvisnosti, upodobljene na sliki[16] 33 in v preglednici 33.

SLIKA 33: PRIKAZ ODVISNOSTI S POMOČJO ALGORITMA NAIVE BAYES



Algoritem Naive Bayes računa odvisnosti v obe smeri. V mojem primeru je računal na primer P(natančnost|podatkovni tipi) = P(natančnost) * P(podatkovni tipi|natančnost) / P(podatkovni tipi). Slika, prenesena neposredno iz *MS Analysis Services*, prikazuje domnevo, da lastnosti FIRPM (podatkovni tipi) vplivajo na kakovost podatkov, nasprotno pa ne. Zavedati se je treba omejitev opravljenega izračuna, saj je število statističnih enot majhno, zato tudi nisem mogel zagotoviti ustreznega vzorca statističnih enot za preverjanje. Kljub temu pa menim, da je izračun

[16] Izračun je bil narejen v modulu *Analysis Services*, ki je del MS SQL. Uporabljena uporabniška rešitev omogoča tudi slikovni prikaz in je v disertaciji tudi prikazan, izračune pa je težko brati.

pokazal smer odvisnosti ter skupaj s kavzalnim pristopom, modeliranjem in teoretičnimi izhodišči omogoča sprejetje smeri povezav, tj. FIRPM -> kakovost podatkov.

PREGLEDNICA 33: IZRAČUNI Z ALGORITMOM NAIVE BAYES

Odvisna spremenljivka	Vrednost odvisne spremenljivke	Neodvisna spremenljivka	Vrednost neodvisne spremenljivke	Verjetnost dogodka
Natančnost	>= 4,33	Podatkovni tipi	>= 3,96	95,5 %
Doslednost	>= 4,48	Podatkovni tipi	>= 3,96	100 %
Popolnost	>= 4,40	Podatkovni tipi	>= 3,96	100 %
Doslednost	< 1,97	Normalizacija	< 3,08	100 %

V modeliranju je bilo pokazano, katera lastnost FIRPM naj bi vplivala na posamezno razsežnost kakovosti podatkov. Izračun linearne regresije je temeljil na vključitvi vseh lastnosti FIRPM, ki izhajajo iz modeliranja, pri posamezni razsežnosti, potem pa je SPSS sam izključeval neodvisne spremenljivke, ki ne vplivajo na odvisno spremenljivko. Na podlagi R^2 sem za posamezno razsežnost izbral najboljši model.

Natančnost

Izpeljani model pravi, da je natančnost funkcija podatkovnih tipov in deklarativnih omejitev. Izračun koeficientov korelacije pokaže povezanost med spremenljivkama podatkovni tipi in natančnost ter spremenljivkama deklarativne omejitve in natančnost. Povezanost je pozitivna, vendar ni močna. Nasprotno, je bolj šibka. Vendar kljb temu govori, da z večjo uporabo deklarativnih omejitev ter uporabo ustreznih podatkovnih tipov povečamo natančnost. Ker koeficient korelacije nakazuje le povezanost, bi lahko trdil tudi drugače: da večja natančnost vpliva na deklarativne omejitve in podatkovne tipe. Vendar modeliranje in logičnost razmišljanja pokažeta, da je takšna povezava nemogoča oziroma nesmiselna.

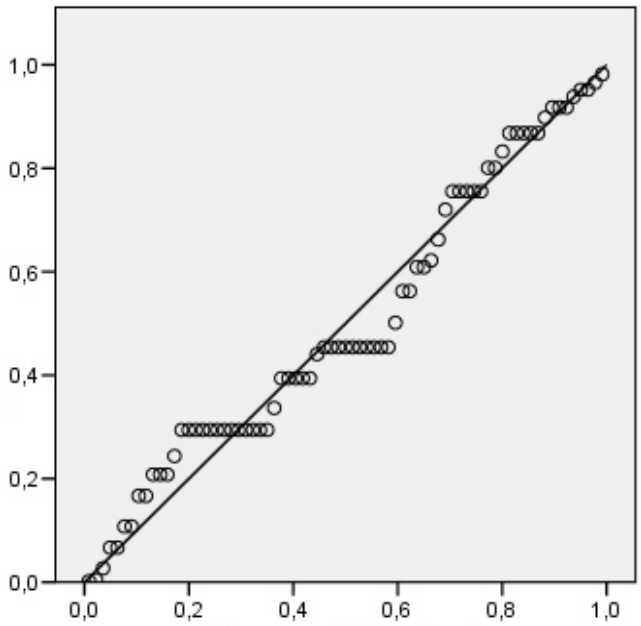
Jakost povezave izraža pomisleke, ki se porajajo pri modeliranju. Podatkovni tipi in deklarativne omejitve imajo omejeno moč pri zagotavljanju natančnosti. Ta moč se nekoliko poveča, kadar ravnamo z »govorečimi« podatki, na primer EMŠO, ali s podatki, ki imajo »vgrajeno« logiko, na primer davčna številka, transakcijski račun. Kot dopolnilo koeficientu korelacije sem izračunal tudi linearno regresijo (preglednica 34), ki pojasnjuje povezavo.

PREGLEDNICA 34: REGRESIJSKA ANALIZA ZA ODVISNO SPREMENLJIVKO NATANČNOST

Popravljen R^2 = 0,141 F = 6,747 (točna stopnja značilnosti 0,003)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	2,292	0,553	4,141	0,000
Podatkovni tipi	0,309	0,114	2,722	0,008
Deklarativne omejitve	0,159	0,071	2,250	0,028
Odvisna spremenljivka: Natančnost				

Popravljen R^2 je za obravnavano področje dokaj nizek, kar pritrjuje omejeni moči deklarativnih omejitev ter podatkovnih tipov pri zagotavljanju natančnosti. R^2 dejansko sporoča, da je treba dejavnike, ki vplivajo na natančnost podatkov, iskati izven podatkovnega modela, na primer v poslovni logiki, uporabniškem vmesniku, načinu dela, organizacijskih predpisih.

SLIKA 34: PRIKAZ VREDNOSTI ZA ODVISNO SPREMENLJIVKO NATANČNOST



Zdaj lahko uporabim v knjigi predstavljeni enačbi, in sicer:

$$D_{21} = a + b_1 * P_1 + b_2 * P_2$$

$$D_{21} = 2{,}292 + 0{,}309 * P_1 + 0{,}159 * P_2,$$

kjer je D_{21} – natančnost; P_1 – podatkovni tipi; P_2 – deklarativne omejitve.

Grafikon na sliki 34 prikazuje, da je linerna regresija ustrezna funkcija za pojasnitev odvisnosti med odvisno spremenljivko natančnost ter neodvisnima spremenljivkama podatkovni tipi in deklarativne omejitve.

Doslednost

Model nakazuje, da je doslednost funkcija podatkovnih tipov, deklarativnih omejitev, normalizacije, glavnih ter tujih ključev. Kot izhaja iz izračunanih koeficientov korelacije je najmočnejša povezava med spremenljivkama »normalizacija« ter »doslednost«. Njuno povezanost sem nakazal že z modeliranjem, statistična analiza pa je povezanost še dodatno potrdila. Povezanost je pozitivna ter srednje močna, kar pomeni, da z večjo normalizacijo dosežemo večjo doslednost.

PREGLEDNICA 35: REGRESIJSKA ANALIZA ZA ODVISNO SPREMENLJIVKO DOSLEDNOST

Popravljen R^2 = 0,313 F = 7,659 (točna stopnja značilnosti 0,000)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	–0,278	0,809	–0,343	0,733
Glavni ključi	0,054	0,109	0,496	0,622
Tuji ključi	0,099	0,060	1,631	0,107
Podatkovni tipi	0,389	0,151	2,575	0,012
Deklarativne omejitve	0,261	0,089	2,952	0,004
Normalizacija	0,303	0,125	2,412	0,019
Odvisna spremenljivka: Doslednost (metoda Enter)				
Popravljen R^2 = 0,323 F = 12,118 (točna stopnja značilnosti 0,000)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	0,020	0,712	0,028	0,933
Podatkovni tipi	0,403	0,155	2,593	0,012
Deklarativne omejitve	0,285	0,088	3,249	0,002
Normalizacija	0,326	0,129	2,522	0,014
Odvisna spremenljivka: Doslednost (metoda Stepwise)				

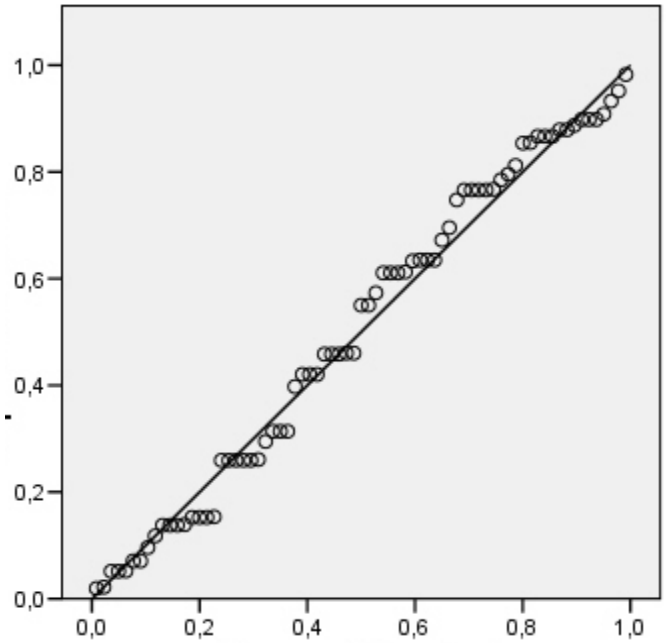
Tudi med podatkovnimi tipi in doslednostjo je pozitivna in srednje močna povezava. Ustreznejše podatkovne tipe glede na namen uporabe uporabljamo, večjo doslednost podatkov v

podatkovnem modelu dosežemo. Podobno je tudi z deklarativnimi omejitvami. Povezanost deklarativnih omejitev z doslednostjo je pozitivna in šibka, vendar še zmeraj statistično značilna. Statistično pa ni značilna povezava med tujimi ključi in doslednostjo, čeprav je vrednost točne stopnje značilnosti blizu 0,05. Takšen rezultat je mogoče pojasniti z vrednostjo spremenljivke »neposredni dostop v zbirko podatkov«. Omenil sem že, da je ta vrednost majhna, kar nakazuje, da združbe v zbirko podatkov neposredno ne posegajo, temveč le posredno z uporabo uporabniške rešitve.

Nižja povprečna vrednost spremenljivke »tuji ključi«, višja povprečna vrednost spremenljivke »neposredni dostop v zbirko podatkov« (2,92) ter precej visoka povprečna vrednost spremenljivke »doslednost« napeljujejo k sklepu, da vlogo tujih ključev v anketiranih združbah opravlja programska koda. Prav tako ni statistično značilna povezava med glavnimi ključi in doslednostjo, kar je mogoče pojasniti z vrednostjo spremenljivke »število polj v glavnem ključu«, ki nakazuje na »suhoparne« glavne ključe.

Tudi konstanta pri linearni regresiji ni statistično značilna, kar nakazuje na dejavnike, ki niso bili zajeti v linearni regresiji in jih bo treba proučiti z nadaljnjim raziskovalnim delom. Konstante ne morem zanemariti, saj lahko doslednost dosežemo tudi s popolnoma neustreznimi podatkovnimi tipi, denormaliziranim podatkovnim modelom ter brez deklarativnih omejitev (na primer ena preglednica izključno s poljem oziroma polji BLOB).

SLIKA 35: PRIKAZ VREDNOSTI ZA ODVISNO SPREMENLJIVKO DOSLEDNOST



Kot dopolnilo koeficientu korelacije izračunam tudi linearno regresijo (preglednica 35), in sicer v prvem delu so v linearni model vnesene vse neodvisne spremenljivke z metodo Enter, v drugem delu pa je bila uporabljena metoda Stepwise, ki je izločila statistično neznačilne neodvisne spremenljivke. Za izpad spremenljivke »glavni ključi« ter »tuji ključi« je najbrž kriv vzorec podatkov, saj je iz dobljenih podatkov mogoče sklepati, da slovenske anketirane združbe namesto tujih ključev v podatkovnih modelih uporabljajo preverjanje v programski kodi, za glavne ključe pa je mogoče sklepati, da so »vsebinsko suhoparni«, na primer (DelavecID in OrganizacijaID).

Popravljen R^2 linearnega modela z metodo Stepwise je 0,323, kar pomeni, da lahko 32 % variance doslednosti pojasnim z linearnim vplivom podatkovnih tipov, deklarativnih omejitev in normalizacije. Iz regresijskega modela je razvidno, da zagotavljanje doslednosti dosežemo z ustreznimi podatkovnimi tipi, deklarativnimi omejitvami in normalizacijo.

$$D_{22} = a + b_1 * P_1 + b * P_2 + b_3 * P_3 + b_4 * P_4 + b_5 * P_5$$

$$D_{22} = 0{,}020 + 0{,}403 * P_1 + 0{,}285 * P_2 + 0{,}326 * P_3,$$

kjer je D_{22} – doslednost, P_1 – podatkovni tipi, P_2 – deklarativne omejitve, P_3 – normalizacija.

Iz slike 35 je razvidno, da je linearna regresija zadosten izračun za pojasnjevanje modela odvisnosti med doslednostjo kot odvisno spremenljivko ter neodvisnimi spremenljivkami podatkovni tipi, deklarativne omejitve in normalizacija.

Popolnost

Model nakazuje, da je popolnost funkcija normalizacije, deklarativnih omejitev in glavnih ključev. Koeficient korelacije pokaže pozitivno povezanost popolnosti in deklarativnih omejitev. Povezava je pozitivna in srednje močna, kar nam sporoča, da lahko z deklarativnimi omejitvami povečamo popolnost. Omenjena lastnost FIRPM je zlasti ustrezna, kadar želimo preprečiti vnos vrednosti NULL oziroma praznih vrednosti. Povezava med spremenljivkama glavni ključi in razsežnostjo popolnost je statistično značilna, vendar šibka, kar je verjetno posledica manjšega števila polj, ki sestavljajo glavni ključ, ter posledica »vsebinske suhoparnosti« omenjenih polj.

Spet sta bila izračunana dva modela, v prvem so bile uporabljene vse neodvisne spremenljivke, ki izhajajo iz modela, v drugem pa samo statistično značilne neodvisne spremenljivke.

Linearna regresija, dobljena s pomočjo metod Enter in Stepwise, je prikazana v preglednici 36. Popravljen R^2 (metoda Stepwise) je 0,277, kar pomeni, da lahko 27,7 % variance popolnosti pojasnimo z linearnim vplivom deklarativnih omejitev in normalizacije. Za iskanje dejavnikov, ki

vplivajo na popolnost, se moramo kot pri natančnosti usmeriti na dejavnike izven podatkovnega modela. Pri metodi Stepwise je izpadla neodvisna spremenljivka »glavni ključi«, kar je nakazoval že koeficient korelacije, saj je stopnja zaupanja blizu vrednosti, ki ločuje statistično značilne podatke od neznačilnih.

Kot pri doslednosti tudi pri popolnosti konstanta ni statistično značilna, čeprav je blizu vrednosti 0,05, kar kaže na dejavnike, ki niso zajeti v linearni regresiji.

Tudi za pojasnjevanje vpliva lastnosti FIRPM na razsežnost popolnost kakovosti podatkov je linearna regresija ustrezen model in ni potrebe po drugačnem modelu (slika 36).

PREGLEDNICA 36: REGRESIJSKA ANALIZA ZA ODVISNO SPREMENLJIVKO POPOLNOST

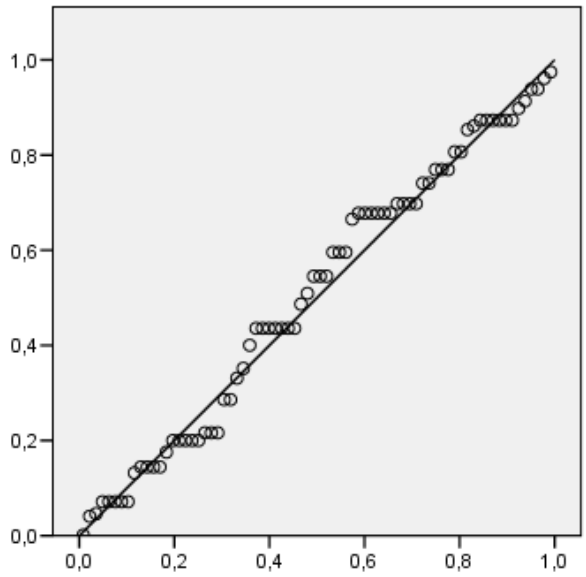
Popravljen R^2 = 0,296 F = 11,221 (točna stopnja značilnosti 0,000)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	0,273	0,680	0,402	0,689
Deklarativne omejitve	0,286	0,091	3,151	0,002
Normalizacija	0,471	0,118	3,979	0,000
Glavni ključi	0,190	0,112	1,700	0,094
Odvisna spremenljivka: popolnost (metoda Enter)				
Popravljen R^2 = 0,277 F = 14,988 (točna stopnja značilnosti 0,000)	**Regresijski koeficient**	**Ocena standardne napake**	**t-test**	**Točna stopnja značilnosti**
Regresijska konstanta	0,970	0,550	1,765	0,082
Deklarativne omejitve	0,310	0,091	3,421	0,001
Normalizacija	0,478	0,120	3,996	0,000
Odvisna spremenljivka: popolnost (metoda Stepwise)				

$$D_{23} = a + b_2 * P_2 + b_3 * P_3 + b_4 * P_4$$

$$D_{33} = -0{,}970 + 0{,}310 * P_2 + 0{,}478 * P_3,$$

kjer je D_{33} – popolnost, P_2 – deklarativne omejitve, P_3 – normalizacija.

SLIKA 36: PRIKAZ VREDNOSTI ZA ODVISNO SPREMENLJIVKO POPOLNOST



Če upoštevamo pesimistično in optimistično oceno pri proučevanih razsežnostih kakovosti podatkov, je kakovost podatkov (enačba (1) in enačba (5)):

$Q = f(K_{2j}) = g(D_{2j}, U_{2j})$, pri čemer je $D_{2j} = k(P_n)$

- Q = ((2,292 + 0,309 * P_1 + 0,159 * P_2) * (0,020 + 0,403 * P_1 + 0,285 * P_2 + 0,326 * P_3) * (0,970 + 0,310 * P_2 + 0,478 * P_3)) – pesimistična ocena;
- Q = min((2,292 + 0,309 * P_1 +0,159 * P_2),(0,020 + 0,403 * P_1 + 0,285 * P_2 + 0,326 * P_3), (0,970 + 0,310 * P_2 + 0,478 * P_3)) – optimistična ocena;

Q – kakovost podatkov, P_1 – podatkovni tipi, P_2 – deklarativne omejitve, P_3 – normalizacija.

9.4.6 Razprava na podlagi statistične analize in povzemanje drugih ugotovitev

S statistično analizo sem se lotil dokazovanja vpliva FIRPM na kakovost podatkov. Pri kakovosti podatkov sem obravnaval štiri razsežnosti, od katerih je bila ena psihološka in posledično ni neposredno odvisna od FIRPM, kakovost zbirke podatkov pa sem meril s petimi lastnostmi. Pri glavnem ključu sem posebej preverjal, ali število polj, ki sestavljajo glavni ključ, vpliva na kakovost podatkov. Rezultati niso presenetljivi, saj so nanje kazale že pretekle lastne delovne izkušnje ter opravljeno modeliranje.

Statistična analiza je pokazala, da so slovenske anketirane združbe precej podobne združbam v ZDA. Čeprav se zavedajo pomembnosti podatkov in jih imajo za pomembno premoženje, z njimi očitno ne ravnajo tako, saj je velik delež anketiranih združb imel težave s kakovostjo podatkov. Prav tako še vedno največji delež pobud za ravnanje s kakovostjo podatkov prihaja od uporabnikov in službe za informatiko. Razlog je precej preprost. Kakovost podatkov je mlado, še dokaj neuveljavljeno področje, ki še ni povsem prodrlo v zavest managementa. Vendar bo, o tem ni dvoma. Sistemi za poslovno obveščanje postajajo čedalje bolj priljubljeni in management so bo moral zavedati, da je kakovost podatkov eden od temeljev uspeha za uspešno dokončanje projektov.

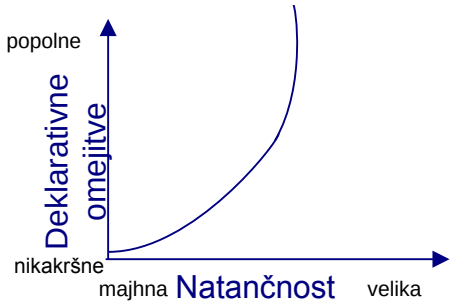
Anketirane združbe očitno upoštevajo namen uporabe, saj si razlike med oceno kakovosti podatkov in težavami zaradi kakovosti podatkov drugače ne morem razlagati. Delež anketiranih združb, ki imajo težave zaradi kakovosti podatkov, je namreč manjši kot delež anketiranih združb, ki imajo težave v kakovosti podatkov. Anketirane združbe imajo manjše ali večje težave s podatki, vendar so ti še vedno ustrezni za nameravano uporabo. Nameravana uporaba se kaže kot osrednji pojem, ki ga mora management združb proučiti, saj ima največji vpliv na merila za ocenjevanje kakovosti podatkov. Od nameravane uporabe je torej odvisno ravnanje s kakovostjo podatkov.

Med proučevanimi razsežnostmi kakovosti podatkov izstopa zaupanje. Zaupanje ni neposredno odvisno od podatkovnega modela, pozitivno pa je odvisno od natančnosti in doslednosti. Omenjena ugotovitev je pomembna, saj kaže na morebitne ukrepe za zagotovitev večjega zaupanja v podatke. Neposredno od podatkovnega modela oziroma zbirke podatkov pa so odvisne ostale proučevane razsežnosti, tj. natančnost, doslednost in popolnost. Statistična analiza je jasno pokazala, da s FIRPM najtežje vplivamo na razsežnost »natančnost«. Zagotovitev delne sintaktične natančnosti je mogoča z ustreznimi podatkovnimi tipi ter deklarativnimi omejitvami, zagotovitev semantične natančnosti pa je pravzaprav nemogoča in je odvisna od dejavnikov, ki s FIRPM niso povezani. Raziskava TDWI (Russom 2006) je pokazala, da med najpogostejše vzroke za nekakovostne podatke spada vnos podatkov zaposlenih. Torej je treba zagotoviti ustrezne organizacijske predpise in njihovo upoštevanje ter izobraževanje zaposlenih, hkrati pa več pozornosti posvetiti načrtovanju PIS. Malenkostna sprememba v uporabniškem vmesniku in zbirki podatkov lahko zelo pozitivno vpliva na kakovost podatkov. Vnos imena in priimka, na primer, v eno polje namesto v ločeni polji lahko ima neprijetne posledice za poznejše analize in obdelave, spet odvisno od nameravane uporabe. Uporaba spustnih seznamov namesto navadnih vnosnih polj lahko zagotovi skoraj popolno sintaktično natančnost, pri zagotavljanju semantične natančnosti pa je njihov vpliv spet omejen. Zaposlenemu ne more nič preprečiti vnosa sintaktično natančnega, a semantično napačnega podatka, zato bi morali pri

občutljivih podatkih uvesti dodatno preverjanje in hkrati vzpostaviti sistem kaznovanja namernega vnosa nenatančnih oziroma širše nekakovostnih podatkov.

Večji vpliv s FIRPM pa imamo na razsežnosti doslednost in popolnost. Levji delež za zagotovitev doslednosti opravimo z ustreznimi deklarativnimi omejitvami. Na popolnost bi lahko vplivali tudi z lastnimi podatkovnimi tipi, kjer bi preprečili vnos praznih vrednosti oziroma vredosti NULL, vendar je ta vpliv po mojem mnenju zanemarljiv. Podatkovni tipi, deklarativne omejitve ter normalizacija so torej osrednji dejavniki, ki vplivajo na proučevane razsežnosti. Med njimi so podatkovni tipi – s tem mislim na izdelavo lastnih podatkovnih tipov – najbrž najtežje razumljivi. Njihova prevelika zapletenost povzroči nerazumevanje in posledično manjšo možnost uporabe. Podatkovni tipi in deklarativne omejitve pa niso edino sredstvo za zagotavljanje doslednosti. Tudi normalizacija je pomemben dejavnik. Če je bila do zdaj normalizacija obravnavana kot sredstvo za zmanjšanje ponavljanja in s tem kot sredstvo za zagotavljanje doslednosti, pa je bilo dokazano, da normalizacija pomembno udejanja semantična pravila, hkrati pa vpliva tudi na popolnost. Statistično nisem mogel dokazati povezanosti tujih ključev in glavnih ključev z doslednostjo podatkov ter glavnih ključev s popolnostjo. Vzrok za neuspeh pri statističnem dokazovanju najverjetneje leži v sestavi vzorca, kar pa je že bilo pojasnjeno.

SLIKA 37: PRIKAZ ODVISNOSTI MED DEKLARATIVNIMI OMEJITVAMI IN DOSLEDNOSTJO



Popolnost je naslednja razsežnost, kjer imajo deklarativne omejitve osrednjo vlogo, predvsem pri zagotavljanju odsotnosti vrednosti NULL. Modeliranje je nakazalo tudi možnost vpliva glavnih ključev na popolnost, vendar statistično povezave nisem uspel dokazati (z linearno regresijo, medtem ko sem jo s koeficientom korelacije uspel dokazati). Tudi v tem primeru je verjetno vzrok v sestavi vzorca. Pri postavljanju predvsem deklarativnih omejitev moramo biti previdni, saj s tem lahko negativno vplivamo na natančnost, njen semantični vidik (slika 37). Omenjeno trditev je preprosto dokazati. Predpostavimo, da računovodski servis izračunava plače za zasebnika. Pogosto se zgodi, da računovodja ne ve, kdaj bo zasebnik nakazal denar, vendar pa deklarativna omejitev od njega zahteva vnos »datum izplačila«. Računovodja je prisiljen vnesti približen datum izplačila, ki je po navadi sicer pravilen, zgodi pa se, da zasebnik ne plača na tisti datum in podatek o datumu izplačila postane nenatančen.

Statistična analiza je pokazala povezavo med FIRPM ter kakovostjo podatkov, vendar je hkrati opozorila, da večina dejavnikov, ki vplivajo na kakovost podatkov, obstaja izven FIRPM, torej v programski kodi, uporabniškem vmesniku, sistemski opremi, zaposlenih, organizacijskih predpisih.

Ravnanja s kakovostjo podatkov se je torej treba lotiti na vseh ravneh.

9.5 Potrditev glavne teze

V knjigi sem uporabil tri načine proučevanja izbranega problemskega področja:

- proučitev literature,
- modeliranje,
- statistična obdelava podatkov.

Proučena literatura je nakazala smer, v kateri se je skrivala rešitev obravnavanega raziskovalnega problema. Raziskovanje se je nadaljevalo v modeliranju in končalo s statistično obdelavo.

Rezultati raziskave so prikazani v preglednici 37. Čeprav dveh izvedenih trditev nisem mogel potrditi s statistično analizo (linearno regresijo), modeliranje vseeno nakazuje vpliv. Vzrok za neuspeh je najverjetneje v sestavi vzorca.

Končna ugotovitev se torej glasi: Fizična izvedba relacijskega podatkovnega modela **vpliva** na kakovost podatkov v PIS.

PREGLEDNICA 37: PREGLEDNICA REZULTATOV RAZISKAVE

Podatkovni model vpliva na kakovost podatkov v PIS.	**Modeliranje**	**Potrditev s statistično analizo**
Glavni ključi podatkovnih entitet vplivajo na kakovost podatkov v PIS.	√	*
Tuji ključi podatkovnih entitet vplivajo na kakovost podatkov v PIS.	√	!
Podatkovni tipi polj vplivajo na kakovost podatkov v PIS.	√	√
Deklarativne omejitve polj vplivajo na kakovost podatkov v PIS.	√	√
Normalizacija v zbirki podatkov vpliva na kakovost podatkov v PIS.	√	√

Legenda: * – potrditev s koeficientom korelacije in zavrnitev z linearno regresijo.

10 SKLEP

Fizična izvedba relacijskega podatkovnega modela vpliva na kakovost podatkov v PIS. Raziskava je to nedvoumno dokazala. Toda dokazala je še več. Čeprav fizična izvedba relacijskega podatkovnega modela vpliva na kakovost podatkov, je ta vpliv omejen. Večina dejavnikov, ki vplivajo na kakovost podatkov, je nepovezana s podatkovnim modelom in je v sestavi uporabniške rešitve, združbi ter poslovnem okolju. Samo tiste združbe, ki se bodo posvetile vsem dejavnikov, bodo imele možnost, da ustrezno ravnajo s kakovostjo podatkov in jo tudi zagotovijo.

Kako se bodo združbe lotile obvladovanja kakovosti, je odvisno od namena uporabe. Namen uporabe je ključen za vse ravni ravnanja s kakovostjo podatkov. Logično je, da bodo morale združbe, katerih osrednji poslovni učinek so podatki (banke, zavarovalnice in druge finančne inštitucije), kakovosti podatkov posvetiti več pozornosti kot druge združbe.

Cilj knjige je pokazati odvisnosti med podatkovnim modelom in kakovostjo podatkov. To ji je tudi uspelo. Pri raziskovanju obravnavane tematike pa obstaja še nekaj nerešenih vprašanj oziroma dilem.

Največkrat se je pojavila dilema v razmerju med opredelitvijo razsežnosti »natančnost« in »doslednost«. Zdi se, da »doslednost« v tistem delu, ko govori o spoštovanju semantičnih pravil, posega na področje semantične natančnosti. Pojem semantično pravilo je namreč zelo ohlapen in prav ta ohlapnost nas mnogokrat zapelje v področje razsežnosti »natančnost«. V tem pogledu bi bilo smiselno razmisliti o natančnosti kot podmnožici doslednosti. Nenatančen podatek je lahko tudi dosleden, medtem ko nedosleden podatek ne more biti natančen. Razjasnjevanje omenjenega razmerja ponuja obilico priložnosti za nadaljnje raziskovanje.

Linearna regresija je nakazala, da pri popolnosti in doslednosti obstajajo še nekateri dejavniki, ki jih bo treba raziskati. Koeficienta pri linearni regresiji namreč nista statistično značilna. V določenih primerih se takšen koeficient lahko zanemari oziroma statistična neznačilnost ni pomembna, vendar bi bilo treba za sprejetje takšne odločitve podrobneje raziskati ti dve razsežnosti.

Pri obdelavi statistične značilnosti so se mi porajale tudi nove ideje, ki bi jih bilo treba raziskati. Ena od njih je podrobnejša proučitev Orrove trditve, da se kakovost podatkov poslabša s starostjo PIS. Z nadaljnjimi raziskavami želim proučiti, kako (če) uvedba skladišča podatkov upočasni proces slabšanja kakovosti podatkov.

Druga pomembna dilema, ki ponuja izziv, pa je objektivna ocena FIRPM. Na področju objektivnega ocenjevanja kakovosti podatkov je bilo storjenega že veliko, čeprav se določene zamisli porajajo tudi na tem področju. Področje objektivnega ocenjevanja podatkovnega modela pa je skoraj popolnoma neraziskano. Pri pisanju knjige so se pojavile določene zamisli objektivnih mer, na primer razmerje med stolpci, ki niso ustrezni glede na vrsto podatkov, shranjenih v teh stolpcih, in vsemi stolpci; razmerje med številom manjkajočih tujih ključev in vsemi tujimi ključi.

Izziv pomeni tudi dopolnitev enotne opredelitve kakovosti podatkov, ki je bila razvita v knjigi z upoštevanjem verjetnostnih porazdelitev.

Naj se knjiga konča z naslednjo mislijo: Prihaja obdobje poglobljenega ravnanja s kakovostjo podatkov, v katerem bo morala večina pobud za ravnanje s kakovostjo podatkov priti od managementa združb. To bo pomenilo, da se je management začel zavedati, da je velik del uspeha poslovanja združb odvisen od kakovosti podatkov.

STVARNO KAZALO

algoritem, 72, 154, 155
analitik kakovosti, 6, 39, 64, 79, 100
analiza podatkov, 83, 96
check constraint, 50, 52, 104, 107, 108, 109, 110, 118, 120, 121, 124, 126, 128, 132, 134, 136, 137, 138, 140, 142, 156, 157, 158, 160, 161, 162, 164
čiščenje podatkov, 71, 72, 73, 74, 75, 87
  ročno, 75
dobavitelji, 63, 69, 87
entiteta, 107
Entiteta, 7, 10, 44, 51, 67, 74, 92, 104, 105, 110, 118, 134
ETL, 66, 68, 69, 70, 178
funkcija, 36, 97, 139, 156, 158, 160
glavni ključ, 8, 62, 78, 81, 82, 83, 106, 108, 109, 110, 119, 134, 135, 138, 140, 151, 160, 161, 162
Herrensteinov izrek, 95
hibridni model, 21
hierarhični model, 23
informacija, 13, 36, 43, 54, 87, 88, 90, 93, 94, 95
informacijska revolucija, 3
informacijska tehnologija, 2, 3, 9, 64, 91, 95, 96, 176
informacijski sistem, 2, 35, 53, 54, 58, 62, 63, 65, 89
intuitivno-čustven način, 17
ISACA, 91, 92, 175
kakovost podatkov, 3, 4, 6, 7, 8, 9, 10, 11, 29, 33, 34, 35, 37, 38, 39, 40, 44, 45, 46, 47, 48, 56, 60, 64, 66, 69, 72, 74, 77, 78, 79, 85, 97, 98, 99, 100, 102, 103, 105, 106, 107, 109, 110, 111, 113, 116, 117, 120, 125, 126, 132, 134, 135, 136, 139, 140, 141, 143, 148, 149, 150, 152, 153, 154, 155, 162, 163, 165, 167
kupci, 2, 30, 35, 41, 42, 43, 44, 63, 86, 95, 96, 99, 127
management, 2, 5, 19, 29, 31, 37, 75, 78, 85, 86, 94, 117, 146, 147, 163, 168, 172, 173, 174, 175, 176, 177
manjkajoče vrednosti, 72, 73, 118
mrežni model, 22, 23, 26
Naive Bayes, 154, 155, 156
normalizacija, 104, 105, 109, 110, 118, 134, 136, 138, 139, 140, 142, 158, 160, 161, 162, 164
objekt, 13, 14, 22, 23, 25, 75, 94
objektivno ocenjevanje, 78
objektni model, 25, 26
odločitve, 1, 13, 14, 15, 16, 17, 18, 19, 30, 31, 32, 35, 37, 54, 56, 75, 94, 96, 97, 116, 117, 134, 147, 148, 167
olap, 5, 37, 38, 63, 116
organizacija, 1, 2, 3, 4, 5, 10, 19, 29, 30, 32, 33, 35, 36, 38, 39, 43, 44, 49, 57, 58, 60, 62, 63, 64, 66, 67, 70, 71, 74, 78, 81, 85, 86, 88, 89, 91, 92, 95, 96, 98, 100, 103, 113, 114, 115, 116, 117, 118, 132, 135, 136, 140, 141, 142, 143, 144, 145, 147, 148, 149,150, 152, 159, 160, 163, 167
panoga, 3, 144
podatek, 5, 6, 8, 9, 14, 19, 21, 23, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 43, 47, 48, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 64, 65, 66, 67, 68, 69, 71, 72, 77, 78, 80, 81, 82, 83, 84, 86, 87, 89, 90, 91, 93, 94, 95, 96, 97, 99, 100, 101, 102,107, 108, 111, 113, 114, 115, 116, 117, 120, 121, 123, 134, 137, 146, 148, 149, 153, 156, 163, 164, 167
podatkovni model, 7, 11, 21, 23, 25, 26, 36, 38, 49, 51, 63, 70, 83, 89, 90, 104, 105, 108, 114, 115, 116, 118, 127, 132, 133, 134, 136, 140, 151, 152, 154
podatkovni tip, 10, 26, 82, 104, 108, 109, 110, 115, 118, 121, 128, 131, 136, 137, 138, 140, 142, 152, 155, 156, 157, 158, 160, 162, 164
pogajalska moč kupcev, 3
prerez podatkov, 61, 77
  podatkovna pravila, 81, 84
  strukturna analiza, 80
racionalno-analitičen način, 17
raziskava, 6, 34, 64, 111, 112, 114, 142
razsežnost kakovosti podatkov
  doslednost, 45, 46, 49, 50, 51, 59, 79, 102, 103, 104, 110, 114, 118, 124, 125, 126, 129, 130, 134, 135, 136, 137, 138, 139, 140, 142, 148, 149, 152, 158, 159, 160, 163, 164, 167
    semantična doslednost, 50
    strukturna doslednost, 50

dostopnost, 46, 47, 57, 58, 70, 94, 103
jedrnatost, 46
natančnost, 29, 35, 45, 46, 47, 48, 49, 53, 54, 55, 58, 59, 67, 68, 79, 81, 95, 103, 104, 110, 114, 118, 120, 130, 132, 133, 136, 137, 138, 139, 140, 142, 148, 149, 152, 155, 156, 157, 158, 163, 164, 167
semantična natančnost, 47
sintaktična natančnost, 47
popolnost, 45, 46, 51, 52, 59, 79, 103, 104, 105, 108, 110, 114, 120, 134, 135, 136, 138, 139, 140, 142, 147, 148, 149, 152, 160, 161, 162, 163, 164
pravočasnost, 35, 45, 46, 47, 55, 56, 58, 59, 94, 103, 104
ustreznost, 45, 46, 47, 56, 94, 98, 103, 122, 137, 151
varnost, 46, 61
zaupanje, 34, 46, 53, 54, 58, 64, 97, 103, 142, 147, 148, 149, 163
relacijski podatkovni model, 6, 7, 9, 11, 26, 50, 51, 81, 100, 104, 109, 110, 112, 114, 118, 125, 126, 134, 135, 138, 139, 142, 143, 152, 154, 155, 156, 160, 161, 162, 163, 164, 165, 168
rudarjenje po podatkih, 6, 36, 39, 40, 60, 69, 71, 73, 83, 171, 172, 175, 177
Semiotika, 93, 94
sistemi za upravljanje baz podatkov, 9, 10, 23, 25, 26, 50, 61, 63, 81, 104, 107, 108, 109, 112, 114, 115, 116, 118, 130, 134, 137, 140, 151, 152, 154
relacijskih baz podatkov, 1
strategija, 41, 91
strateški management, 2, 5, 31
stroški nekakovostnih podatkov, 29, 91
taktični management, 31, 37
TIQM, 9, 86, 87, 90
upravljanje odnosov s kupci, 41, 171, 172, 174, 176
vedenjsko-preudaren način, 17
vnos podatkov, 5, 64, 65, 66, 110, 146, 163
zaznavanje, 6, 43, 74
zbiranje in posredovanje podatkov, 60
znanje, 3, 6, 11, 14, 16, 19, 38, 43, 62, 64, 70, 81, 100, 113, 114

LITERATURA

Adelman, Sid. 2006. *Ask the experts*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1062746&topicId=230005 (6. 6. 2007)

Adler, Steven. 2006. *Data Governance*. http://www.ibm.com/itsolutions/datagovernance (17.12. 2007)

Ambler, Scott. 2003. *Agile database techniques: effective strategies for the agile software developer*. Indianapolis: Wiley.

———. 2006. *Whence Data Quality?* , http://www.ddj.com/dept/database/196900212?cid=Ambysoft (15. 7. 2007)

Ambler, Scott in Pramod Sadalage. 2006. *Refactoring databases: evolutionary database design*. Upper Saddle River: Addison Wesley.

Awad, Elias in Malcolm Gotterer. 1992. *Database Management* Boyd & Fraser Pub. Co.

Bailey, James in Sammy Pearson. 1983. Development of a Tool for Measuring and Analyzing Computer User Satisfaction. *Management Science* 29 (5): 530-546.

Ballou, Donald in Harold Pazer. 1982. The impact of inspector fallibility on the inspection policy serial production system. *Management Science* 28 (4): 387-399.

———. 1985. Modeling data and process quality multi-input multi-output. information systems. *Management Science* 2: 150-162.

———. 1987. Cost/quality tradeoffs for control procedures information systems. *Management Science* 15 (6): 509-521.

———. 2003. Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts. *IEEE Transactions on Knowledge and Data Engineering* 15 (1): 240-243.

Ballou, Donald, Richard Wang, Harold Pazer in Giri Kumar Tayi. 1998. Modeling information manufactoring systems to determine information quality product. *Management Science* 44 (4): 462-484.

Baroni, Aline, Fernando Abreu in Colar Calero. 2005. *Finding where to apply object-relational database schema refactorings: an ontology-guided approach*. http://ctp.di.fct.unl.pt/QUASAR/Resources/Papers/2005/BaroniJISBD05.pdf (11. 12. 2007)

Batini, Carlo, Enrico Nardelli in Roberto Tamassia. 1986. A Layout algorithm for data flow diagrams. *IEEE Transactions on Software Engineering* 12 (4): 538-546.

Batini, Carlo in Monica Scannapieca. 2006. *Data quality: concepts, methodologies and techniques*. Berlin: Springer.

Bauer, Kent. 2004. *The Power of Metrics:KPIs: Not All Metrics are Created Equal*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1014522&topicId=1029230 (29. 8. 2007)

Bellis, Mary. 2006. *Inventors of the Modern Computer*. http://inventors.about.com/od/mstartinventions/a/microprocessor.htm (15. 7. 2007)

Berson, Alex, Stephen Smith in Kurt Thearling. 1999. *Building data mining applications for CRM*. New York: McGraw-Hill.

Bertin, Lou. 2004. *The Observer: The Bliss Of Consistency* http://www.informationweek.com/showArticle.jhtml?articleID=20900061 (24. 8. 2007)

Birk, Donna. 2007. *Decision-Making Truth and Consequences*. http://www.selfgrowth.com/articles/Birk5.html (8. 8. 2007)

Bowen, Paul, David Fuhrer in Frank Guess. 1998. Continuously Improving Data Quality in Persistent Databases *Data Quality* 4 (1).

Brackstone, Gordone. 1999. *Managing Data Quality in a Statistical Agency*. http://dsbb.imf.org/vgn/images/pdfs/scpap.pdf (13. 7. 2007)

Burns, Larry. 2005. *The Ugly Truth about Data Quality*. http://www.dmreview.com/article_sub.cfm?articleId=1028545 (15. 5. 2007)

Cats-Baril, William in Roland Thompson. 2002. *Information Technology and Management*. Irwin: McGraw-Hill.

Cerf, Christopher in Victor S. Navasky. 1998. *The Experts Speak : The Definitive Compendium of Authoritative Misinformation* Villard.

Chapman, Arthur. 2005. *Principles of Data Quality* http://www.gbif.org/prog/digit/data_quality/DataQuality.pdf (28. 7. 2007)

Chung, Charles. 2002. *The Top Five Obstacles to Achieving Data Quality*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=5106&topicId=230005 (6. 9. 2007)

Chung, WooYoung, Craigh Fisher in Richard Wang. 2002. What skills matter in data quality? Paper read at 7th International Conference on Information Quality, 8-9 November 2002.

Clarke, Roger. 1994. *Path of Development of Strategic Information Systems Theory*. http://www.anu.edu..au/people/Roger.Clarke/SOS/StratISTh (14. 7. 2007)

Codd, Edgar. 1975. Implementation of Relational Data Base Management Systems. *FDT - Bulletin of ACM SIGMOD* 7 (3).

———. 1975. Understanding Relations. *FDT - Bulletin of ACM SIGMOD* 7 (1).

———. 1979. Extending the Database Relational Model to Capture More Meaning. *ACM Transactions on Database Systems* 4 (4).

———. 1990. *The relational model for database management* Reading: Addison-Wesley.

Connor, Tim. 2007. *Decisions And Consequences*. http://eziniearticles.com/?Decisions-And-Consequences&id=319064 (20. 7. 2007)

Cook, Rick. 1997. *Is a hybrid database in your future?* . http://sunsite.uakom.sk/sunworldonline/swol-02-1997/swol-02-objects.html (11. 8. 2007)

CRMAdvocate. B. l. *Defining Customer Relationship Management (CRM)* 2007. http://www.realmarket.com/crmdefine.html (6. 9. 2007)

Dasu, Tampraparni, Gregg Vesonder in Jon Wright. 2003. Data quality through knowledge engineering. Paper read at Conference on Knowledge Discovery in Data 24-27 Avgust 2003.

Dasu, Tamraparni in Theodore Johnson. 2003. *Exploratory data mining and data cleaning*. New York: John Wiley.

Date, Christopher. 1995. *An introduction to database systems*, *Addison-Wesley systems programming series.* Reading: Addison-Wesley Pub. Co.

Delone, W.H. in E. R. McLean. 1992. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research* 3 (1): 60-95.

Deming, W. Edwards. 1986. *Out of the crisis*. Cambridge: Massachusetts Institute of Technology.

Dongre, Kuldeep. 2004. *Data Cleansing Strategies*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1012952&topicId=230005 (11. 8. 2007)

Donohue, Eric , Tony Chang in Jon Bostwick. 2004. *Clean Up Your Data*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1015666&topicId=230005 (10. 6. 2007)

Dravis, Frank. 2003. *A Simple Case of ROI*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=6477&topicId=230005 (10. 6. 2007)

Elmasri, Ramez in Sham Navathe. 2000. *Fundamentals of database systems*. Reading: Addison-Wesley.

English, Larry. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. New York: Wiley.

———. 2002. *The Essentials of Information Quality Management*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=5690&topicId=230005 (6. 7. 2006)

———. 2006. *Information Quality and Reference Data*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1048494&topicId=230005 (5. 8. 2007)

Eppler, Martin in Markus Helfert. 2004. A Clasification and Analysis of Data Quality Costs. Paper read at International Conference on Information Quality 11-15 Oktober 2004, at Miami, Florida, USA.

Feltham, Jerald. 1968. The Value of Information. *The Accounting Review* 43 (4): 684-696.

Ferengul, Corey. 2006. *Error In, Error Out: Safeguarding the Quality of the Data Warehouse*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1059121&topicId=230005 (10. 6. 2006)

Fisher, Craigh, Eitel Lauria in Carolyn Matheus. 2007. In Search of an Accuracy Metric. Paper read at International Conference on Information Quality 9-11 November 2007 at Cambridge, USA.

Flere, Sergej. 2000. *Sociološka metodologija*. Maribor: Pedagoška fakulteta.

Fox, C., A Levitin in Thomas Redman. 1994. The notion of data and its quality dimensions. *Information processing & management* 30 (1): 9-19.

Frazer, Gordon. 2007. *The Tech Lab*. http://news.bbc.co.uk/1/hi/technology/6981704.stm (7. 1. 2008)

Friedman, Paul. 2006. *Eliminate "Garbage In, Garbage Out": Build a Better Foundation for High-Quality Customer Data*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1053116&topicId=230005 (12. 4. 2006)

Fuchs, Gabriel. 2002. *Practical Tips to Analyze Your Data Quality*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=6055&topicId=230005 (10. 6. 2006)

Gable, Guy. 1994. Integrating Case Study and Survey Research Methods: An Example in Information Systems. *European Journal of Information Systems* 3 (2): 112-126.

Geiger, Jonathan. 2004. *Data Quality Management: The Most Critical Initiative You Can Implement*. http://www2.sas.com/proceedings/sugi29/098-29.pdf (6. 9. 2007)

Gilfillan, Ian. 2002. *Introduction to Relational Databases*. http://www.databasejournal.com/sqletc/article.php/1469521 (10. 12. 2007)

Goasdoué, Virginie , Sylvaine Nugier, Dominique Duquennoy in Brigitte Laboisse. 2007. An Evaluation Framework For Data Quality Tools. Paper read at International Conference on Information Quality 9-11 November 2007 at Cambridge, USA.

Government, BC. B. l. *BC Government Information Resource Management Glossary* 2008. http://www.cio.gov.bc.ca/other/daf/IRM_Glossary.asp#d (13. 12. 2007)

Grassi, Roberto. 2007. *Backup Strategies*. http://www.grsoftware.net/backup/articles/backup_strategies.html (8. 9. 2007)

Greenhalgh, Trisha in Rod Taylor. 1997. *How to read a paper: Papers that go beyond numbers (qualitative research)*. http://www.bmj.com/cgi/content/full/315/7110/740 (9. 1. 2008)

Greenstein, Marilyn in Todd M. Feinman. 2000. *Electronic Commerce: Security, Risk Management and Control*. Boston: Irwin McGraw-Hill.

Grossman, Robert , Simon Kasif, Reagan Moore, David Rocke in Jeff Ullman. 1998. Data Mining Research: Opportunities and Challenges. A Report of three NSF Workshops on Mining Large, Massive and Distributed Data. Illinois: University of Illinois.

Groznik, Aleš in Dejan Vičič. 2006. Menedžment portfelja projektov službe za informatiko. *Uporabna informatika* 14 (4): 219-225.

———. 2007. Menedžment poslovnih procesov in operativnih tveganj. *Uporabna informatika* 15 (2): 65-69.

Gunasekaran, Angappa, Omar Khalil in Syed Mahbubur Rahman. 2003. *Knowledge and information technology management : human and social perspectives*. Hershey: Idea Group Pub.

Hall, Curtis. 2005. *The importance of data quality for data warehousing and CRM* http://searchcrm.techtarget.com/originalContent/0,289142,sid11_gci1059931,00.html (21. 7. 2007)

Halloran, Denis, Susan Manchester, John Moriarty, Robert Riley, James Rohrman in Thomas Skramstad. 1978. Systems Development Quality Control *MIS Quarterly* 2 (4): 1-13.

Han, Jiawei in Micheline Kamber. 2001. *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.

Hand, Steve in Jane Chandler. 1998. *INTRODUCTION TO OBJECT-ORIENTED DATABASES*. http://www.odbms.org/download/005.01%20Chandler%20Introduction%20to%20Object-Oriented%20Databases%20September%201998.pdf (20. 7. 2007)

Harris, Robert. 1998. *Decision Making Techniques*. http://www.virtualsalt.com/crebook6.htm (20. 7. 2007)

Harrison, E. Frank. 1975. *The managerial decision-making process*. Boston: Houghton Mifflin.

Harvey, Jerry B. 1988. *The Abilene paradox and other meditations on management*. Lexington: Lexington Books.

Hay, David. 2003. *Analyze That? Ensuring Data Quality for Successful CRM*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=7606&topicId=230005 (10. 6. 2006)

Haynie, Mark. 1981. The relational/network Hybrid data model for Design Automation Databases. Paper read at Conference on Design Automation at Nashville, Tennessee, USA.

Hipp, Jochen, Ulrich Guntzer in Gholamreza Nakhaeizadeh. 2000. Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations* 2 (1): 58–64.

Hipp, Jochen, Ulrich Guntzer in Udo Grimmer. 2001. Data Quality Mining -Making a Virtue of Necessity. Paper read at ACM SIGMOD

Huang, Kuan-Tsae, Yang Lee in Richard Wang. 1999. *Quality information and knowledge*. Upper Saddle River: Prentice Hall.

Hudicka, Joseph. 2003. *Develop a Data Quality Strategy Before Implementing a Data Warehouse*. http://www.dmreview.com/portals/portalarticle.cfm?articleId= 6478&topicId=230005 (10. 6. 2007)

Hussain, Tauqeer, Shafay Shamail in Mian Awais. 2004. Schema transformation - a quality perspective. Paper read at Multitopic Conference, 24-26 December 2004, at Lahore, Pakistan.

IAIDQ. B. l. *International Association for Information and Data Quality* 2007. http://www.iaidq.org/main/about.shtml (29. 12. 2007)

Imai, Masaaki. 1986. *Kaizen (Ky'zen): the key to Japanese competitive success*. New York: Random House Business Division.

———. 1997. *Gemba kaizen: a commonsense low-cost approach to management*. New York: McGraw-Hill.

Inmon, William H., J. D. Welch in Katherine L. Glassey. 1997. *Managing the data warehouse*. New York: Wiley.

International, INFORMATION IMPACT. B. l. *Business Excellence through Information Excellence* 2007. http://www.infoimpact.com/tiqmmethodology.cfm (6. 9. 2007)

ISACA. B. l. *ISACA Overview and History* 2007. http://www.isaca.org/Content/NavigationMenu/About_ISACA/Overview_and_History/Overview_and_History.htm (17. 11.. 2007)

Jarke, Matthias, Manfred Jeusfeld, Christoph Quix in Panos Vassiliadis. 1999. Architecture and Quality for Data Warehouses: An Extended Repository Approach. *Information Systems* 24 (3): 229-253.

JDIQ. B. l. *Journal od Data and Information Quality* 2007. http://jdiq.acm.org/index.htm (11. 12. 2007)

Karacsony, Ken. 2006. *Proactive Data Quality*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1048865&topicId=230005 (12. 4. 2006)

Kennedy, Ruby L. 1997. *Solving data mining problems through pattern recognition*. Upper Saddle River, N.J.: Prentice Hall PTR.

Klein, Barbara. 1998. Data Quality in the Practice of Consumer Product Management:Evidence From the Field. *Data Quality* 4 (1).

Kobeja, Boris. 2001. *Priročnik za pisce strokovih besedil*. Koper: Fakulteta za management.

Kotler, Philip. 1994. *Marketing management : analysis, planning, implementation, and control*. 8th ed, *Prentice-Hall series in marketing*. Englewood Cliffs: Prentice-Hall.

Kwan, Steve. 2006. *MINISIS Database Architecture*. http://www.minisisinc.com/docs/architecture.pdf (17. 6. 2007)

Laudon, Kenneth. 1986. Data quality and due process in large interorganizational record systems. *Communications of the ACM* 29 (1): 4-11.

Lee, Yang, Diane Strong, Beverly Kahn in Richard Wang. 2002. AIMQ: a methodology for information quality assessment *Information&Management* 40 (2): 133-146.

Lee, Yang W. 2006. *Journey to data quality*. Cambridge, Mass.: MIT Press.

Lesjak, Dušan. 1988. Objektivni in subjektivni dejavniki odločanja. *Organizacija in kadri* 21 (3-4): 269-278.

Lesjak, Dušan, Viktorija Sulčič, Cene Bavec, Srečko Natek, Leo Zornada, Uroš Godnov, Boris Šušmak, Aleš Tankosič in Tamara Bertok. 2006. Gradivo za predmet Poslovna informatika. Koper: Fakulteta za Management.

Lincoln, Guisnel, S. Nugier in V. Stéphan. 2004. Contrôle et amélioration de la qualité: Rapport de recherche EDF.

Litwin, Paul. 2004. *Stop SQL Injection Attacks Before They Stop You*. http://msdn.microsoft.com/msdnmag/issues/04/09/SQLInjection/ (6. 9. 2007)

Livesley, Richard. 2006. *Information into action*. http://www.isaca.e-symposium.com/archive250907.php (10. 10. 2007)

Loshin, David. 2001. *Enterprise knowledge management : the data quality approach*. San Diego: Morgan Kaufmann.

———. 2004. *Issues and Opportunities in Data Quality Management Coordination*. http://www.dmreview.com/article_sub.cfm?articleId=1000836 (5. 7. 2007)

———. 2005. *Developing Information Quality Metrics*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1026061&topicId=230005 (10. 6. 2006)

———. 2006. *The Data Quality Business Case:Projecting Return on Investment*. http://i.i.com.com/cnwk.1d/html/itp/informatica_Data_Quality_Business_Case.pdf

Luebberes, Dominik, Udo Grimmer in Matthias Jarke. 2003. Systematic Development of Data Mining-Based Data Quality Tools. Paper read at VLDB.

Madnick, Stuart in Richard Wang. 1992. Introduction to the TDQM research program.

Maletic, Jonathan in Andrian Marcus. 2000. *Data Cleansing: Beyond Integrity Analysis1*. http://citeseer.ist.psu.edu/cache/papers/cs/16282/http:zSzzSzwww.msci.memphis.eduzSz~maleticjzSzpaperszSzIQ2000.pdf/maletic00data.pdf (7. 8. 2007)

Manning, Ian. 1999. *Data Warehousing*. http://www.manning.demon.co.uk/dw.htm (7. 8. 2007)

Markus, Lynne in Cornelis Tunis. 2000. The Enterprise System Experience - From Adoption to Success. In *Framing the domains of IT management : projecting the future-- through the past*, edited by R. W. Zmud. Cincinnati, Ohio: Pinnaflex Education Resources, Inc.

Marshall, Brian. B. l. *Data Quality and Data Profiling* 2007. http://www.telusplanet.net/public/bmarshal/dataqual.htm#DATA_QUALITY_PROFILING (29. julija 2007)

Maydanchik, Arkady. 2007. *Data Quality Assessment*. New Jersey: Technics Publications.

McCue, Andy. 2006. *Poor-quality data is biggest CIO headache*. http://news.zdnet.co.uk/itmanagement/0,1000000308,39267945,00.htm (9. 5. 2006)

Mihelčič, Miran. 1972. Organizacijski pomen informacijskega sistema. Magistrsko delo, Univerza v Ljubljani, Ljubljana.

———. 2003. *Organizacija in ravnateljevanje*. Ljubljana: Fakulteta za računalništvo in informatiko.

Mingers, John. 2001. Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research* 12 (3): 240-259.

Miriyala, Shravan 2007. *Problem Solved:The Need for Data Profiling in Customer Data Integration*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1075086&topicId=1029230 (29. 8. 2007)

MIT. B. l. *International Conference on Information Quality* 2007. http://mitiq.mit.edu/iciq/ (11. 12. 2007)

Monahan, James. 2006. *Essentials of an Effective Backup Strategy*. http://www.redsofts.com/articles/read/158/6771/Essentials_of_an_Effective_Backup_Strategy.html (5. 5. 2007)

Morey, Richard. 1982. Estimating and Improving the Quality of Information in a MIS. *Communications of the ACM* 25 (5): 337-342.

Mullen, Nancy. 2003. *Information for Innovation: Less Bitter: Improving the Quality of Your Data*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=6194&topicId=230005 (10. 6. 2007)

Müller, Heiko in Johann-Christoph Freytag. 2003. *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf (8. 8. 2007)

Naumann, Felix. 2001. Quality-driven Query Planning Berlin University, Berlin.

Naumann, Felix, Johann Freytag in Ulf Leser. 2004. Completeness of Integrated Information Sources. *InformationSystems* 24 (7): 583-615.

Naumann, Felix in Mary Roth. 2004. Information quality: How good are off-the-shelf DBMS? Paper read at International Conference on Information Quality 11-15 Oktober 2004, at Miami, Florida, USA.

Neal, Richard. 2001. *Customer Profitability&CRM*. http://www.crm-forum.com/library/art/art-128/art-128.html (15. 10. 2001)

Neely, Pamela. 2005. *Data Quality Tools for Data Warehousing - A Small Sample Survey*. http://www.ctg.albany.edu/publications/reports/data_quality_tools (17. 5. 2007)

Novo, Jim. 2004. *Drilling Down: Turning Customer Data Into Profits With A Spreadsheet*. Saint Petersburg: Booklocker.com.

Oliveira, Paulo, Fatima Rodrigues in Pedro Henriques. 2006. A Formal Definition of Data Quality Problems. Paper read at International Conference on Information Quality, 19-23 September 2005, at Houston, Texas, USA.

Olle, T. William. 1978. *The Codasyl approach to data base management*. New York: Wiley.
Olson, Jack E. 2003. *Data quality the accuracy dimension*. San Francisco: Morgan Kaufmann Publishers.
Orr, Ken. 1998. Data quality and systems theory *Communications of the ACM* 41 (2): 66-71.
Piattini, Mario, Colar Calero in Marcela Genero. 2001. Table Oriented Metrics for Relational Databases *Software Quality Journal* 9 (2): 79-97.
Piattini, Mario, Colar Calero, Houari Sahraoui in Hakim Lounis. 2006. *Object-relational database metrics*. http://www.iro.umontreal.ca/~sahraouh/papers/lobjet00_1.pdf (15. 7. 2007)
Pierce, Elizabeth. 2003. Pursuing a Career in Information Quality: The Job of the Data Quality Analyst. Paper read at International Conference on Information Quality, 7-9 November 2003.
Pipino, Leo, Yang Lee in Richard Wang. 2002. Data Quality Assessment. *Communications of the ACM* 45 (4).
Postma, Paul. 1999. *The new marketing era : marketing to the imagination in a technology-driven world*. New York: McGraw-Hill.
Pradhan, Shekhar. 2005. Believability as an Information Quality Dimension. Paper read at International Conference on Information Quality, 4-6 November 2005, at MIT, Cambridge, USA.
Pregibon, Daryl. 1997. Data Mining. *Statistical Computing and Graphics* 8 (1).
Pučko, Danijel. 1996. *Strateško upravljanje*. Ljubljana: Ekonomska fakulteta.
Pyle, Dorian. 2003. *Business modeling and data mining*. Boston: Morgan Kaufmann Publishers.
Quality, Statistics Canada. B. l. *Data Quality*. 11. avgust 2006. http://www.statcan.ca/english/edu/power/ch3/quality/quality.htm (28. 7. 2007)
Rahm, Erhard in Hong-Hai Do. 2000. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering* 23 (4).
Raman, Vijayshankar in Joseph M. Hellerstein. 2001. Potter'sWheel: An Interactive Data Cleaning System. Paper read at VLDB, at Rome, Italy.
Ranyard, Rob, Ray Crozier in Ola Svenson. 1997. *Decision Making: Cognitive Models and Explanations* Routledge.
Redman, Thomas. 1992. *Data quality : management and technology*. New York: Bantam Books.
———. 1996. *Data quality for the information age*. Boston: Artech House.
———. 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM* 41 (2).
———. 2001. *Data quality the field guide*. Boston: Digital Press.
———. 2004. Barriers to successful Data Quality Management. *Studies in Communication Sciences* 4 (2):53-68.
———. 2004. *Data: An Unfolding Quality Disaster*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1007211&topicId=230005 (10. 6. 2006)
———. 2005. Measuring Data Accuracy: A Framework and Review *Information Quality* 1.
Rhind, Graham. 2002. *The Trouble with Standards*. http://www.dmreview.com/article_sub.cfm?articleID=5556&topicId=230005 (6. 7. 2006)
Rittman, Mark. 2007. *An Introduction to Real-Time Data Integration*. http://www.oracle.com/technology/pub/articles/rittman-odi.html (19. 11. 2007)
Rud, Olivia. 2001. *Data mining cookbook : modeling data for marketing, risk and customer relationship management*. New York: Wiley.
Rudra, Amit in Emilie Yeo. 1999. Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organisations in Australia. Paper read at International Conference on System Sciences, at Hawaii, USA.

Rusell, Kay. 2006. *Backup Strategies*. http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=112246 (9. 8. 2007)

Russom, Pilip. 2006. *Liability and Leverage - A Case for Data Quality*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1060128&topicId=230005 (30. oktober 2006)

Sadhegi, Akbar in Richard Clayton. 2002. *The Quality vs. Timeliness Tradeoffs in the BLS ES-202 Administrative Statistics*. www.fcsm.gov/01papers/Sadeghi.pdf (19. 7. 2007)

Sanna, Filip. 2006. *Discovering Data Quality*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1056199&topicId=230005 (10. 6. 2006)

Scannapieca, Monica, Paolo Missier in Carlo Batini. 2005. *Data Quality at a Glance*. http://www.dis.uniroma1.it/~monscan/ResearchActivity/Articoli/DS2005.pdf (13. 7. 2007)

Schweitzer, Douglas 2004. *Data Security Debacle Fundamentals For Keeping Information Safe* (23), http://www.processor.com/editorial/article.asp?article=articles/p2623/21p23/21p23.asp&searchtype=&WordList=&bJumpTo=True (17. 8. 2007)

Shankaranarayanan, Ganesan, Richard Wang in Mostapha Ziad. 2000. IP-MAP: Representing the Manufacture of an Information Product. Paper read at Conference on Information Quality, 20-22 Oktober 2000, at Boston, USA.

Silvers, Fon. 2006. *Deming, Data Quality and ETL, Part 1:Point 3 - Cease Dependence on Inspection*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1047156&topicId=230005 (12. 5. 2007)

Sim, Susan. 1996. *Automatic Graph Drawing Algorithms*. http://www.ics.uci.edu/~ses/papers/grafdraw.pdf (5. 6. 2007)

Smith, Anne. 2006. *Aks the experts*. http://www.dmreview.com/portals/portalarticle.cfm?articleId=1062746&topicId=230005 (9. 4. 2007)

Steglich, Christian Erich Gerhard. 2003. The framing of decision situations : automatic goal selection and rational goal pursuit, University of Groningen, Groningen.

Strong, Diane in Olga Volkoff. 2004. A Roadmap for Enterprise System Implementation. *Computer* 37 (6): 22-29.

Tammasia, Roberto, Giuseppe Battista in Carlo Batini. 1988. Automatic graph drawing and readability of diagrams. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 18 (1): 61-79.

Tayi, Giri Kumar in Donald Ballou. 1998. Examining Data Quality. *Communications of the ACM* 41 (2): 54-57.

Terry, Dian. 2006. *More Decisions, More Complexity, More Data*. http://www.teradata.com/t/page/157049/index.html (20. 7. 2007)

Tierstein, Leslie. 2004. *A Methodology for Data Cleansing and Conversion*. http://www.wrsystems.com/whitepapers/dbclean.pdf (16. 8. 2007)

Timmermans, Danielle. 1991. Decision Aids for Bounded Rationalists, University of Groningen, Groningen.

Torr, Mark. 2007. *The New Data Integration Landscape*. http://www.sas.com/news/sascom/2006q1/column_tech.html (14. 8. 2007)

UALR. B. l. *Information Guality Graduate Program* 2007. http://ualr.edu/informationquality/ (11. 12. 2007)

University, Dublin City. B. l. *Ballou & Pazer IQ Dissertation Award* 2007. http://www.computing.dcu.ie/research/dataquality/thesis/award.htm (11. 12. 2007)

———. B. l. *Information Quality* 2007. http://www.computing.dcu.ie/research/dataquality/thesis/ (11. 12. 2007)

Vishnev, Alex in Emil Vishnev. 2005. *If Not a Relational Database, then What?* , http://www.dmreview.com/article_sub.cfm?articleID=1031010 (3. 7. 2007)

W3C. B. l. *World Wide Web Consortium* 2007. http://www.w3.org/ (11. 12. 2007)

Wang, Richard. 1998. A Product perspective on Total Data Quality Management. *Communications of the ACM* 41 (2).

Wang, Richard, Thomas Allen, Wesley Harris in Stuart Madnick. 2003. An Information Product Approach for Total Information Awareness. Paper read at IEEE Aerospace Conference.

Wang, Richard, Veda C. Storey in Christopher Firth. 1995. A Framework for Analysis of Data Quality Research. *IEEE Transactions on Kkowledge and Data Engineering* 7 (4).

Wang, Richard in Diane Strong. 1996. Beyond accuracy: What data quality means to data consumers *Journal of Management Information Systems* 12 (4).

Weik, Martin. 1961. The ENIAC Story. *The Journal of the American Ordnance Association*.

Witten, Ian in Eibe Frank. 2000. *Data mining: practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.

Wittgenstein, Ludwig, David Francis Pears in Brian McGuinness. 2001. *Tractatus logico-philosophicus*. London: Routledge.

Xu, Hongijang. 2003. Critical Success Factors for Accounting Information Systems Data Quality, University of Southern Queensland.

Zaiane, Osmar. 1998. *Boyce-Codd Normal Form*. http://www.cs.sfu.ca/CC/354/zaiane/material/notes/Chapter7/node10.html (27. 6. 2007)