Smals
ICT for society

Return on Experience with Data Quality Tools at Smals

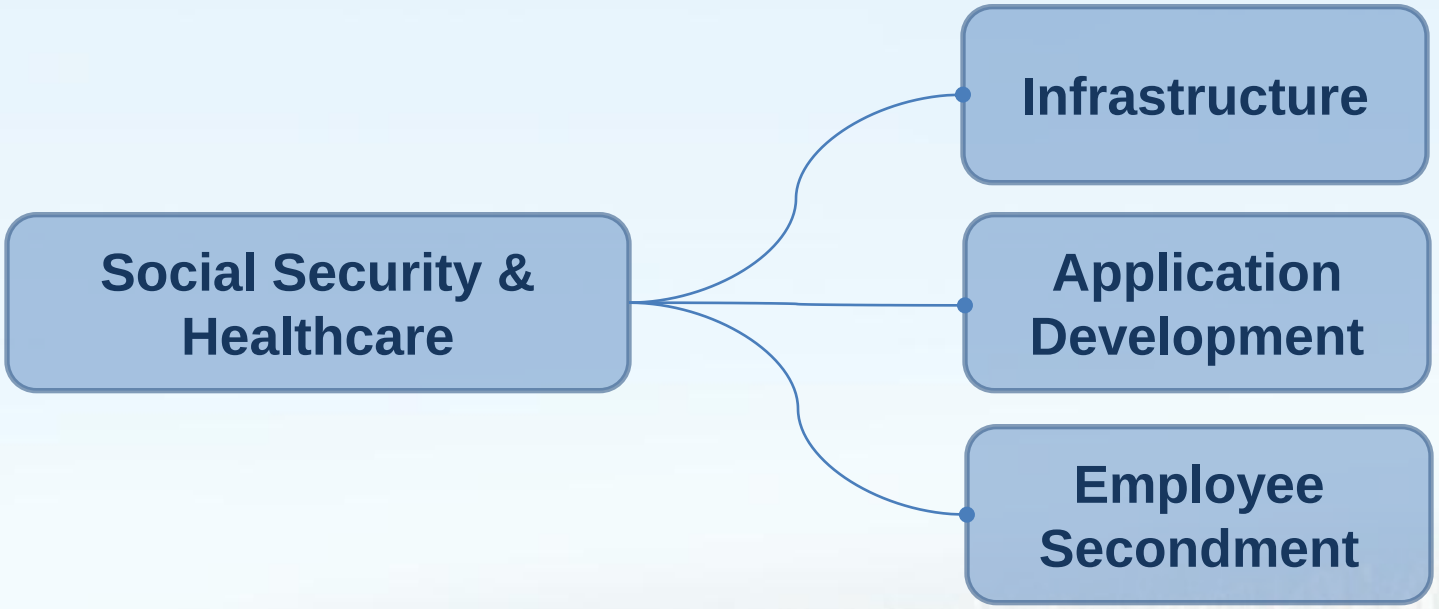
Dries Van Dromme @ULB, 13/03/2013

Contents

- Introduction – Smals
- Introduction – The Reality of Data Quality
- Functionalities of Data Quality Tools
- Organization for Data Governance
- Conclusions

Smals ICT for society

Full IT service with large development components for new applications



ICT for work, health & family life

Smals: one of the largest IT organizations in Belgium

Smals: one of the largest IT organizations in Belgium

top EMPLOYERS 2013 BELGIE/BELGIQUE
CERTIFIED BY THE CRF INSTITUTE

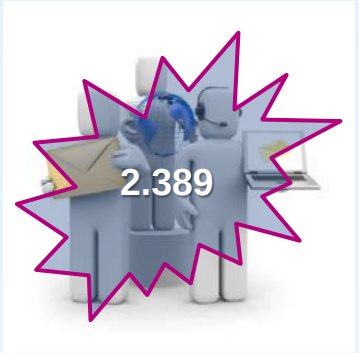
Smals ICT for society

Facts & Figures

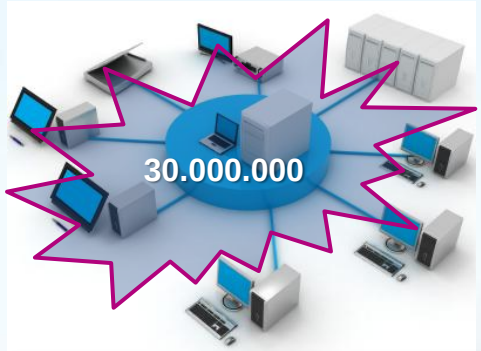
Datacenters

Batch

Applications



Service Management



Network

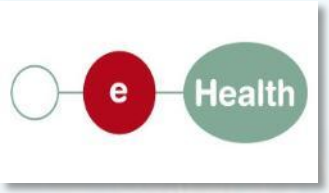
Servers

Storage

Databases

Middleware & Web Services

Some of our members



Some of our projects

vesta

vitalink

gegevens delen, van vitaal belang

Dimona

binc

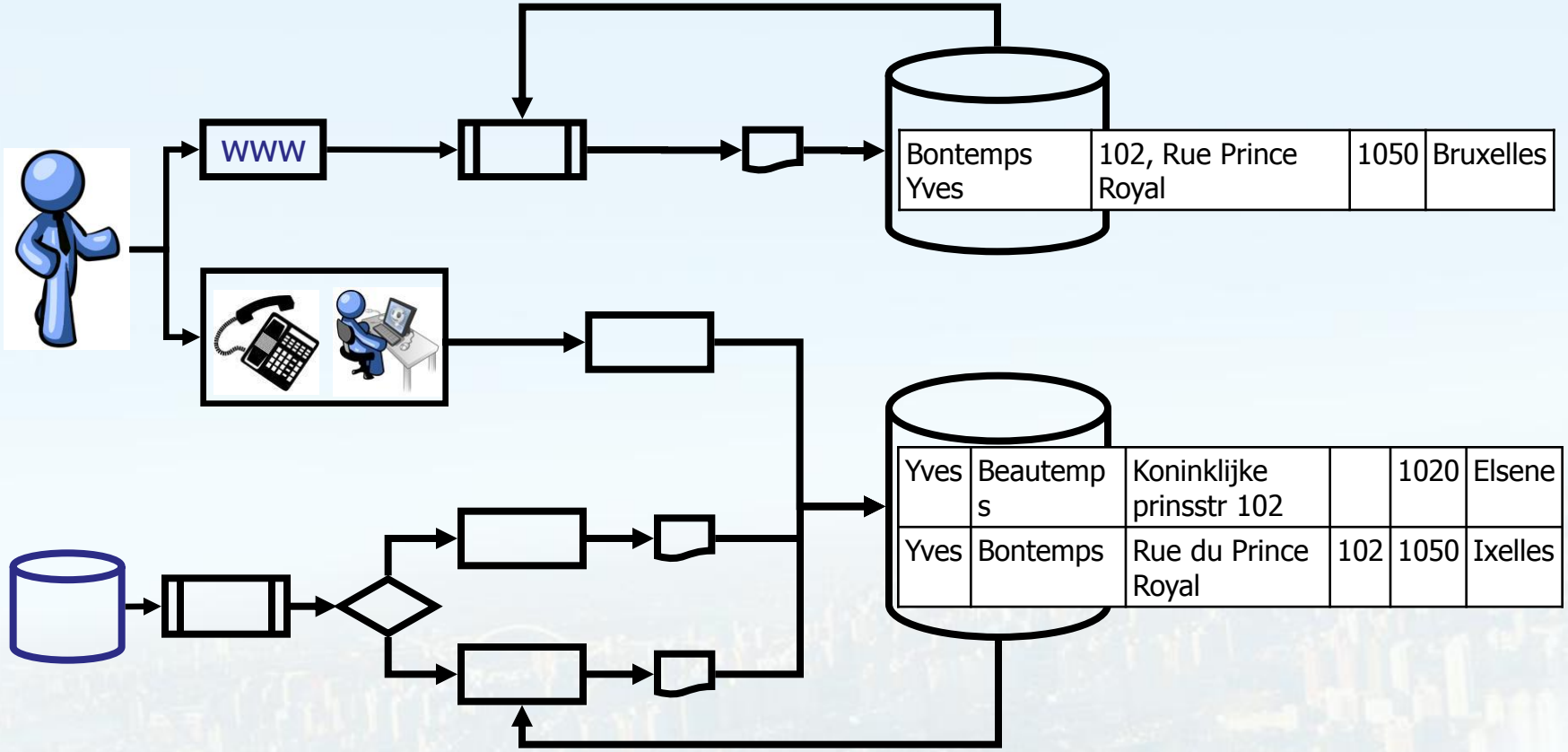
e Health

socialezekerheid.be

Contents

- Introduction – Smals
- Introduction – The Reality of Data Quality
- Functionalities of Data Quality Tools
- Organization for Data Governance
- Conclusions

The Reality of Data Quality (Problems)



Contents

- Introduction – Smals
- Introduction – The Reality of Data Quality
- Functionalities of Data Quality Tools
- Organization for Data Governance
- Conclusions

Functionalities of Data Quality Tools

- Data Profiling
- Data Standaardisatie
- Data Matching / fuzzy matching / record linkage

Real world examples

- Name- and address-cleansing, address validation
- Incoherence detection, fuzzy matching between DB's

Data Profiling
Definition

- Formele audit van gegevens en documentatie
  - *Jack E. Olson, "Data Quality – The Accuracy Dimension"*
  - input: werkelijke data + documentatie, metadata (kwaliteit onbekend)
  - output: Data Quality Issues (kwaliteit in kaart gebracht) + ondersteuning correctie documentatie
- Data Profiling met Data Quality Tools
  - Veld per veld: Column Property Analysis
  - Structuur: referentiële constraints
  - Consistency: business rules

Data Profiling – Column Property Analyse, drill-down (1)

Entities | Analysis | Findings

unt: 6)	Statistic
Source_secondaire(60)	Rows=246496
Metadata	Keys=0 Deps=0
Attributes	Count=16
Reeksnummer(1)	Distribution=100.000
Identificatienummer(2)	Distribution=99.763
Atypemp(3)	Distribution=0.014
Anatjur(4)	Distribution=0.041
Ara(5)	Distribution=0.002
Arem(6)	Distribution=0.001
Adataffil(7)	Distribution=3.876
Adatsup(8)	Distribution=0
Adenomemp(9)	Distribution=98.672
Aadresemp(10)	Distribution=82.704
Aboxemp(11)	Distribution=0.148
Apostemp(12)	Distribution=0
Type	real
Schema Name	Apostemp
Compliance%	100.000%
Inferred type	Integer (Unkn
Unique Values	1155 (0.469%
Patterns	2
Min	0
Max	9992
Min Len	1
Max Len	4
Masks	2
Integers	1155 (100.00.
Dependencies(Discovered)	0 (2)
Attribute Created Date	2008/09/08 1
Acomemp(13)	Distribution=2.401

- (1)-(16) geanalyseerd tijdens load
  - gegenereerde metadata
    - Min, Max, Lengtes,
    - Null Values, Unique Values
    - Patronen,
    - Distributies,
    - Business Rules, …
- Apostemp(12): postcode werkgever
  - vb. patronen, a.h.v. Masks

Data Profiling – Column Property Analyse, drill-down (2)

Entities | Analysis | Findings

unt: 6)	Statistic
Source_secondaire(60)	Rows=246496
Metadata	Keys=0 Deps=0
Attributes	Count=16
Reeksnummer(1)	Distribution=100.000
Identificatienummer(2)	Distribution=99.763
Atypemp(3)	Distribution=0.014
Anatjur(4)	Distribution=0.041
Ara(5)	Distribution=0.002
Arem(6)	Distribution=0.001
Adataffil(7)	Distribution=3.876
Adatsup(8)	Distribution=0
Adenomemp(9)	Distribution=98.672
Aadresemp(10)	Distribution=82.704
Aboxemp(11)	Distribution=0.148
Apostemp(12)	Distribution=0
Type	real
Schema Name	Apostemp
Compliance%	100.000%
Inferred type	Integer (Unkn
Unique Values	1155 (0.469%
Patterns	2
Min	0
Max	9992
Min Len	1
Max Len	4
Masks	2
Integers	1155 (100.00.
Dependencies(Discovered)	0 (2)
Attribute Created Date	2008/09/08 1
Acomemp(13)	Distribution=2.401

Unique Masks

Unique Masks
Attribute = Source_secondaire(60).Apostemp

Mask	Mask Pattern	Value Count	Frequency	Dist %
NNNN	N4	1154	239983	97.358
N	N1	1	6513	2.642

Quality Project | Unique Masks

Unique Values

Unique Values
Attribute = Source_secondaire(60).Apostemp

Value	Frequency	Dist %	Length	Soundex	Metaphor
0	6513	2.642	1		

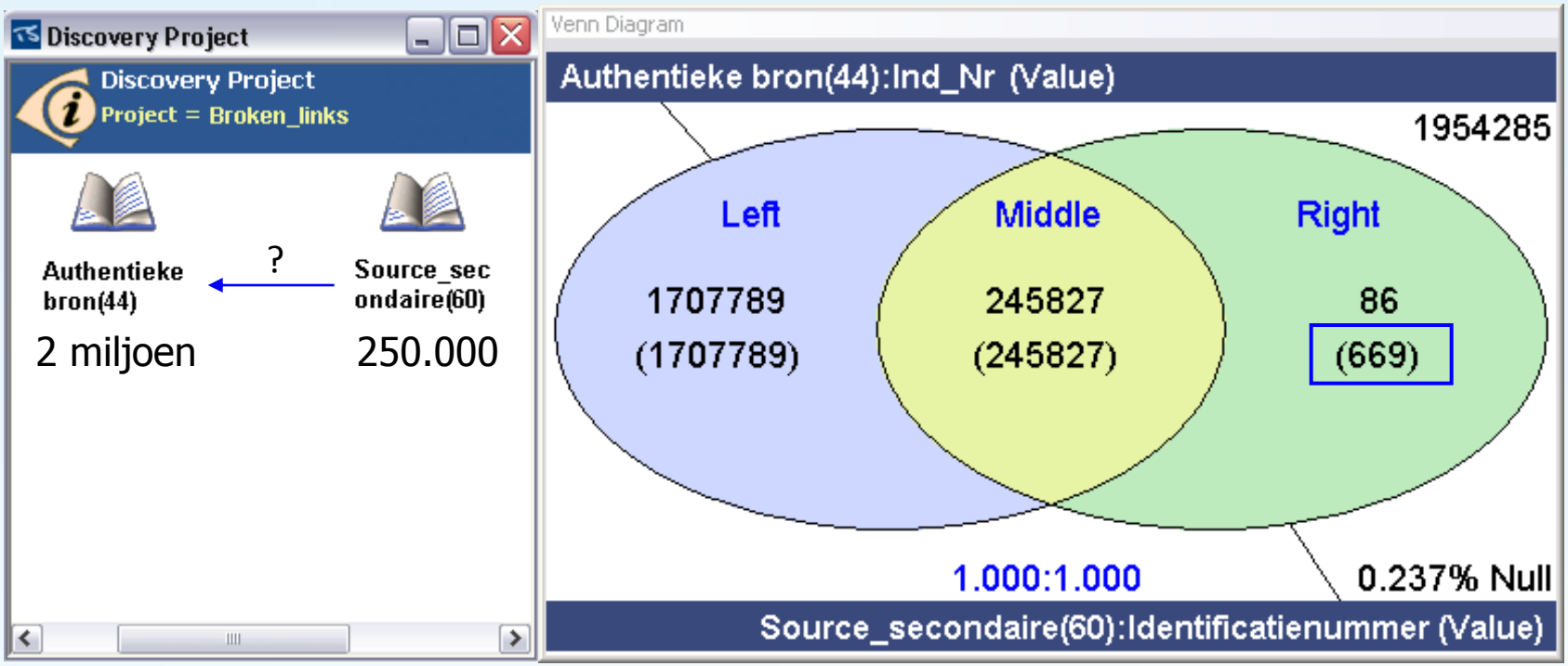
Quality Project | Unique Values

Data rows filtered by selected values in 'Apostemp'

Row	Aadresemp	Apostemp	Acomemp
51	Z.ART.COMMERC. LANGWIES	0	JUNGLINSTER LU
77	STEPHENSONSTRAAT 5	0	JA TIEL NL
90	TORENLAAN 10	0	2215RW VOORHOUT NL
137	HAGENWEG 5A	0	4131 LX VIANEN NL
147	BOEKENDERWEG 18	0	THORN NL
156	OVERGOO 2 C	0	LEIDSCHENDAM NL
206	ALLEE DE LONGCHAMP 2	0	92150 SURESNES FR
220	VAN KOERSVELDWEG 1	0	7665SG ALBERGEN NL
296	PUERTA DEL SOL 7	0	MADRID ES
302	RUE CALMETTE 47	0	VENDIN LE VIEIL FR
335	RUDONK 4	0	BREDA NL
481	RUE DU STADE	0	SAINT FULGENT FR

Smals ICT for society

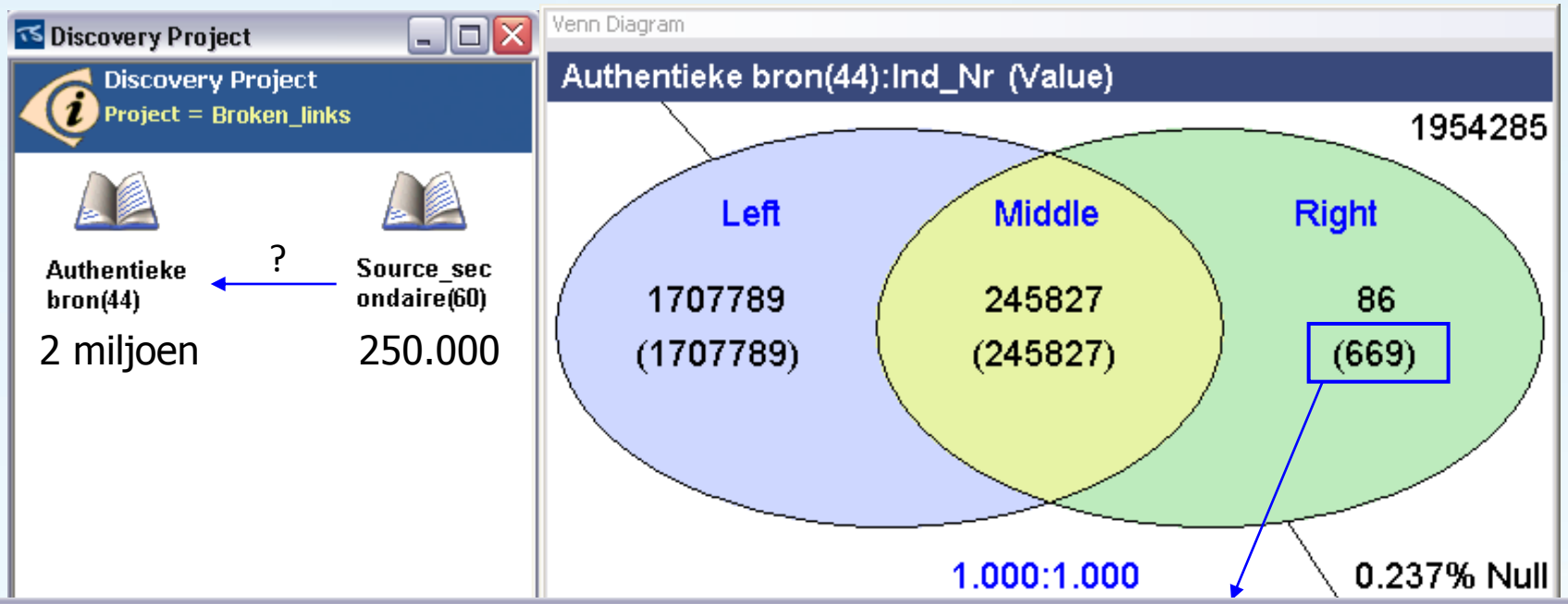
Data Profiling – Referentiële integriteit, foreign keys (1)



- bron 1: Authentieke bron(44)
  - sleutel: Ind_nr
- bron 2: Source_secondaire(60)
  - Identificatienummer verwijst naar Ind_nr's

- **Confrontatie bron 1 – bron 2**
  - 86 waarden niet in authentieke bron
  - in 669 records (er zijn dus dubbels)

Data Profiling – Referentiële integriteit, foreign keys (2)



Non-Joining Right-Hand Rows

Non-Joining Right-Hand Rows
Join = Authentieke bron(44):Ind_nr > - < Source_secondaire(60):Identificatienummer

Row	Reeksnum...	Identificatien...	Atypemp	Anatjur	Adataffil	Adatsup	Adenomemp	Aadresemp	Apostemp	Acomemp
192648	45018	01032	161		1050701		KERKFABRIEK O.L. VROUW GEBOORTE EN SINT ...	SCHELDELAAN 1	8580	AVELGEM
110617	48768	01097	141	2011	1070701		RESIDENTIE DUINHOEK I & II VME	DUINHOEKSTRAAT 123	8660	DE PANNE
201530	35477	05395	111		1060116		ROUKINE ANN	RUE DES PIVOINES 11	1020	BRUXELLES
126830	43906	08998	121	4011	1030801		BV BVBA MICHIELSEN & WITTENS GEASS NOT ...	HANDELSLEI 102	2980	ZOERSEL
239099	49646	08997	111		1080609		BAYON ISABELLE	HOOIENDONKSTRAAT 52	2801	MECHELEN
246142	50537	08996	121	4011	1080616		DAUWEN MARC, LIPSCHUTZ LAURA, DRAULANS ...	GEVAERTLAAN 180	2260	WESTERLO
246146	51428	08993	151		1080519		DEMFOOD BV	VLAARDINGWEG 51	0	3044 CJ ROTTERDAM NL
241041	57168	08994	131	21	1080707		GEMA BOUW BVBA	HOEVESTRAAT 33 B	1755	GOOIK
66445	08990	62118	161		1080401		KERKFABRIEK VAN HET HEILIG HART	HEILIG HARTPLEIN 1	9040	GENT
240701	67363	18310	141	2011	1080701		RESIDENTIE DE BERGEYCK VME	CORTEWALLEDREEF ZN	9120	BEVEREN
246421	70432	13371	131	22	1080701		MICHAEL GERIN SPRL	RUE DU PLAT RIE 73	7390	QUAREGNON
239321	76964	18305	111		1080801		VOGELS ROEL	SCHEPEN DEJONGHSTRA...	3800	ST TRUIDEN
51780	77459	17132	111		1080701		FERLIN JAN	AARTRIJKESTRAAT 15	8480	ICHTEGEM

Data Profiling – Functional Dependencies

Dependencies (Discovered)
Entity Adres Communicatie(350)

Detectie, tijdens load

quality | sample 10.000 | conflicten

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verifi
C Taalcode,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
C Taalregime,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
D Begindatum,C Postcode	C Nis Gemeentecode	Discovered	No	60	98.760	9876	62	129	
D Begindatum,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.990	9999	1	2	
D Ts Lwijz	D Ts Creatie	Discovered	No	60	98.450	9845	109	232	
Gemeentenaam,Landnaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
Straatnaam 21	Straatnaam Voll	Discovered	No	60	99.350	9935	29	65	
Straatnaam Voll	Straatnaam 21	Discovered	No	60	99.140	9914	68	136	

Dependencies (Verified)
Entity Adres Communicatie(350)

Verificatie, on demand

exhaustief, 2.9 miljoen

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verified Date	Verified By
Landnaam	C Landcode	Permanent	Yes	-	99.960	2935363	5	10	2010/04/06 16:17:19	smals

drill-down naar
overzicht van conflicten indien
Quality < 100%

Dependency Conflicts
Adres Communicatie(350) Dependency Landnaam -> {C Landcode}

Frequency	Landnaam	C Landcode
2854		150
233		
1292	Allemagne	173
945	Allemagne	134
3	Angola	341
1	Angola	381
20	Bahamas	425
5	Bahamas	484
18	Tchécoslovaquie	130
5	Tchécoslovaquie	171

- **Drill down to Matching Rows**
- Resolve Conflicts...
- List Corrections...
- Filter...
- Bookmark
- Convert view to Entity...
- Back
- Forward
- Change View
- Export
- Export to Server
- Copy

Smals
ICT for society

Data Profiling – Business Rules (1)

Sub Contractor5(490) Rows=104930
- Metadata Keys=0 Deps=0
  - Attributes 12
  - Rows Loaded 104930
  - Business Rules(Passed) 2(0)
  - Keys(Discovered) 0 (3)
  - Dependencies(Discover... 0 (5)
  - Date Created 2010/07/
  - Entity State Fully Loa
- Attributes Count=12
  - Denomination(1) Distributi
  - Id_1(2) Distributi
  - Id_2(3) Distributi
  - Id_3(4) Distributi
  - Rue(5) Distributi
  - Rue Num(6) Distributi
  - Rue Box(7) Distributi
  - Commune(8) Distributi
  - Postcode(9) Distributi
  - Country Code(10) Distributi
  - Creation Date(11) Distributi
  - Max Work End Date(12) Distributi

Entity Business Rule

Enabled ☑ Name: Présence_Identifiant

Description: Au minimum 1 des 3 types d'identifiant doit être rempli

NOT (Id_1 = "" AND Id_2 = "" AND Id_3 = "")

+ - * / ( ) = <> < <= >= > AND OR NOT IN LIKE

Entity Business Rule

Enabled ☑ Name: Enregistrement à temps

Description: Date de déclaration < date de fin des derniers travaux

[Creation Date] < [Max Work End Date]

+ - * / ( ) = <> < <= >= > AND OR NOT IN LIKE

Choose expression elements from the lists below || |: IF ELSE

Set the threshold to be 100 % of

The test passed on 67.785% of

Save Cancel

Attributes	Bce
Functions	Commune
Literals	Country Code
Operators	Creation Date
	Denomination

Business Rules
Entity = Sub Contractor5(490)

Name	Description	Threshold	Result	Fail Count	Passing Fraction	Pass Count
Enregistrement à temps	Date de déclaration < date de fin des derniers travaux	100	failed	33803	67.785	71127
Présence_Identifiant	Au minimum 1 des 3 types d'identifiant doit être rempli	100	failed	3063	97.081	101867

Smals ICT for society

Data Profiling – Business Rules (2)

Failing Rows (Présence_Identifiant)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
ATMI SARL				R ...	37/A		59300VA...	0	111	2009-05-20 ...	2011-10-21 00:00:0...
ENTREPRIS...				RU...	9		ANDENNE	5300	150	2005-08-18 ...	2006-06-22 00:00:0...
CONCEPT ...				AV...	15		FOREST	1190	150	2004-02-19 ...	2003-12-01 00:00:0...
ABWW SPRL				ZO...	11	n/a	MONT D...	7750	150	2009-05-20 ...	2009-03-31 00:00:0...
UTGES PR...				KL...	6A		OUDEN...	4730 AE	129	2007-09-26 ...	2007-12-31 00:00:0...
MEMOLI B...				G...	37		HASSELT	3511	150	2007-03-20 ...	2011-07-31 00:00:0...
JOSSE JON...				RU...	80		CHARLE...	6031	150	2005-12-02 ...	2006-09-28 00:00:0...
C & S CHA				BE	10		ZOUTLE	3440	150	2006-02-09	2006-07-15 00:00:0

Failing Rows (Enregistrement à temps)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
VAN DIJCK ...	76...			BA...	1		WUUST...	2990	150	2007-10-09 ...	2007-08-31 00:00:0...
COZIER LO...	46...	124...		RU...	n/a	n/a	LIBRAM...	6800	150	2009-05-20 ...	2003-11-28 00:00:0...
CURNET SP...	41...	127...		R ...	63	n/a	ETTERB...	1040	150	2009-05-20 ...	2008-05-31 00:00:0...
DE LENG K...	73...		78297...	TO...	7	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
VAN SCHIJ...	73...		78297...	DA...	2	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
WOOD-B B...	46...	171...		P...	74	n/a	NIEL	2845	150	2009-05-20 ...	2006-04-14 00:00:0...
DESMIDT D...	74...		78242...	DU...	n/a	n/a	ST KATE...	2860	150	2005-06-16 ...	2005-01-31 00:00:0...
RACOMI BV	44	160		G	27	n/a	BEER	3990	150	2009-05-20	2003-05-15 00:00:0

Data Profiling - summary

What's the most difficult part?

DB-schema,
Constraints,
Business Rules,
Documentation, …

Metadata
(inaccurate, incomplete)

Real data
(complete, quality=?)

project
Data Profiling

DQTools

Metadata
(corrected)

Facts about data
(quality = !)

Business people
DQ Analyst team

Data Quality Issues
- lack of standardisation
- schema not respected
- rules not respected

Large % of effort
- getting access
- getting the metadata
- getting the data
- getting the data in the right (normal) form

Data Standardisation
Definition

- Oplossen van gebrek aan standaardisatie
  - binnen één databank
  - overheen verschillende databanken
    - transversaal datamanagement
    - doorbreken van silo's
    - oplossen van inconsistenties in het (her-)gebruik van data-concepten
  - overheen verschillende bedrijven of instellingen
- Interne verwerkingsstap → fuzzy matching
  - best practice
  - matching-resultaten worden betrouwbaarder

Data Standardisation – simple domains

Example: country codes

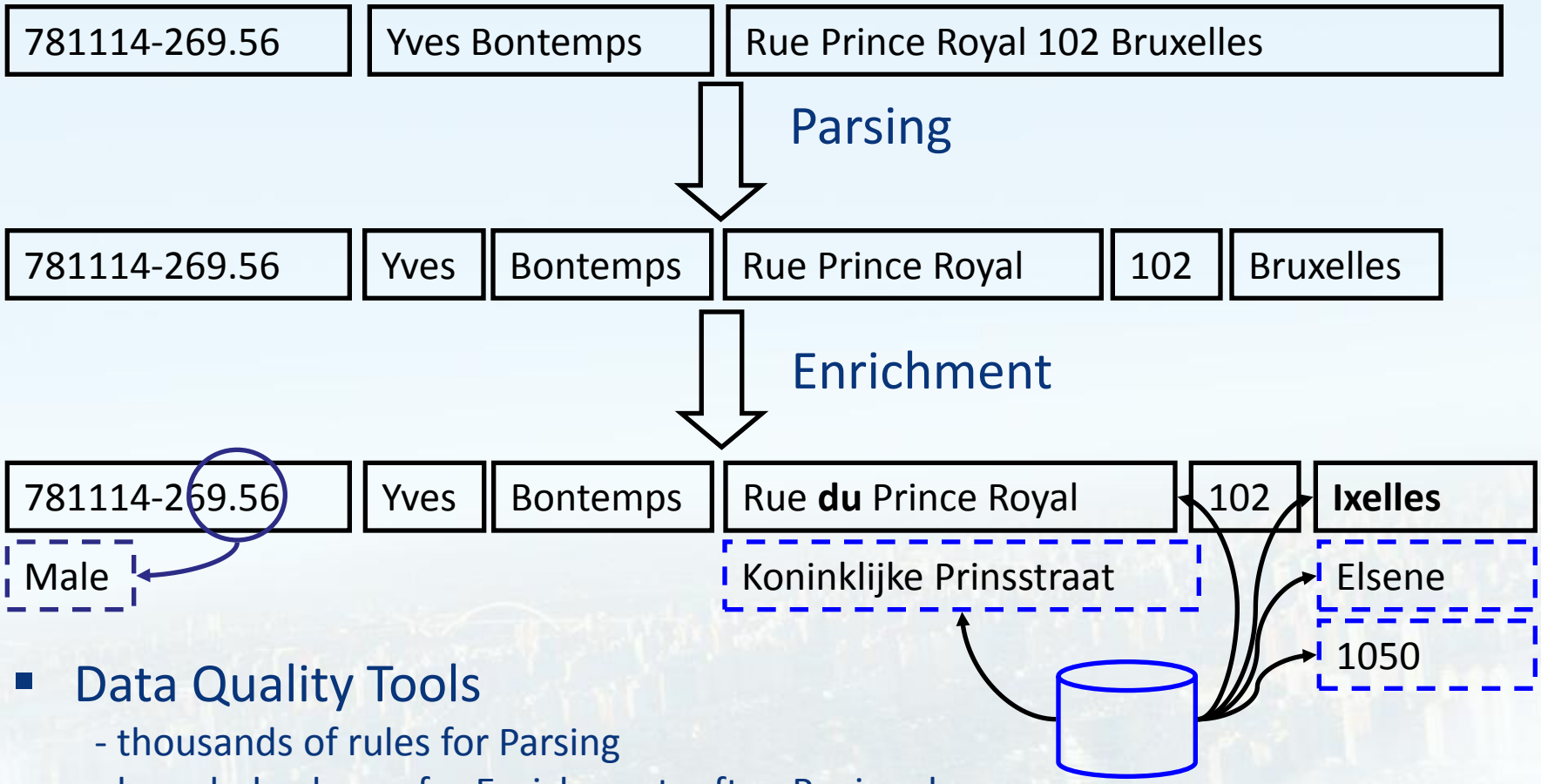
Data Rows (Dynamic)

Entity = pl tranfrmr p26(513)

Tsq Denom	Adres	Boite	Tsq Postcd	Tsq Commune	Orig Landcd	Iso3166 Cd	Iso3166 2l	Iso3166 3l	Iso3166 Nl	Ins Cd	Ins Nl
GREEFS ...					150	056	BE	BEL	België	150	België
JACOBS ...	MAH...			BLACKROC...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
SUIR EN...	OLD...			WATERSFO...	IRL	372	IE	IRL	Ierland		
BONCIK J...	268		067 82	MODRA NAD...	141	703	SK	SVK	Slowakije	141	Slovaakse Republiek
BONCIK J...	HIA...		067 82	MODRA NAD...	SK	703	SK	SVK	Slowakije		
BELROO...	Vizi...		1031	BUDAPEST	115	348	HU	HUN	Hongarije	115	Hongarije (Rep.)
BELROO...	VIZI...		1031	BUDAPEST	H	348	HU	HUN	Hongarije		
VOORBIJ ...	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
VOORBIJ ...	SICI...		1045	AMSTERDAM	NL	528	NL	NLD	Nederland		
DELAT KFT	MAD...		1131	BUDAPEST	H	348	HU	HUN	Hongarije		
DE WAAL...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
FUNDERI...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
SANDMA...	czer...	4	20 349	LUBLIN	122	616	PL	POL	Polen	122	Polen (Rep.)
SANDRA...	CZE...		20 349	LUBLIN	PL	616	PL	POL	Polen		
TPA EDIL ...	Via ...		24129	BERGAMO	128	380	IT	ITA	Italië	128	Italië
TPA GRO...	VIA ...		24129	BERGAMO	I	380	IT	ITA	Italië		
VLASMA...	Stee...		2407 BD	ALPHEN AA...	129	528	NL	NLD	Nederland	129	Nederland
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		

Data Standardisation – complex domains

Example: names and adresses



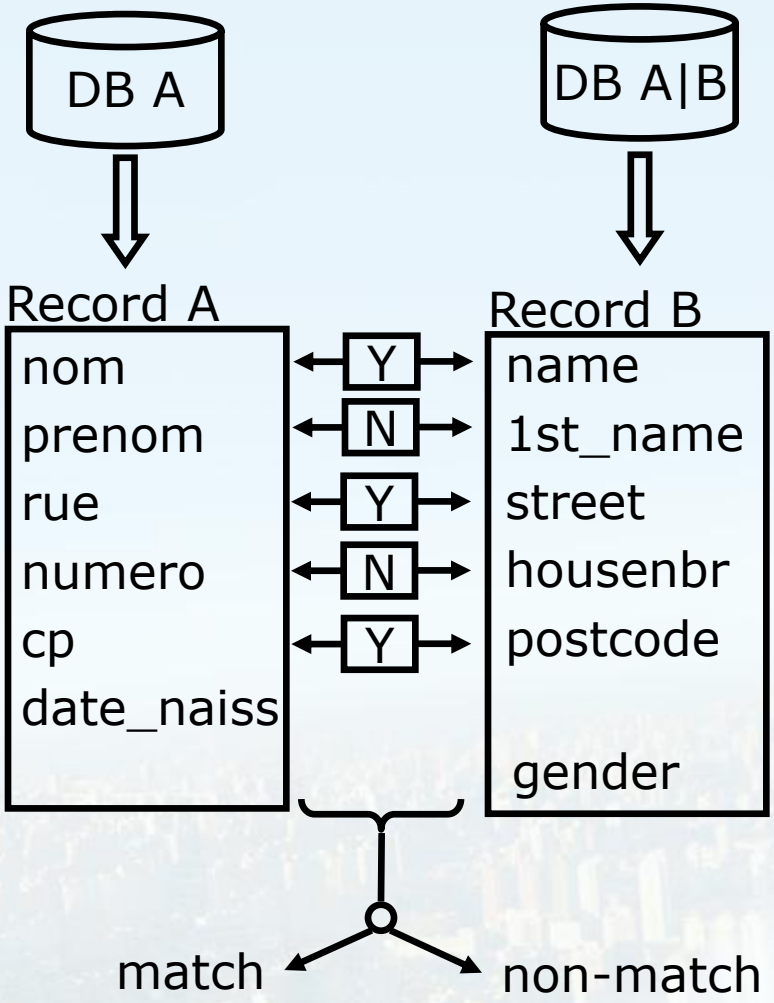
- Data Quality Tools
  - thousands of rules for Parsing
  - knowledge bases for Enrichment, often Regional

Data matching
Definition

- Omgaan met schrijffouten, onnauwkeurigheden, gebrek aan standaardisatie
- om links te kunnen leggen tussen records
  - van één bron (dubbeldetectie)
  - van meerdere bronnen (dubbel- of incoherentie-detectie)
  - ook waar de datamodellen van de te vergelijken bronnen niet overeenstemmen (data integratie)
- met hoge performantie (>=1 miljoen/uur)
- en hoge flexibiliteit
  - definitie vaak initieel onduidelijk
  - wat mag als 'dubbel' of 'incoherent' beschouwd worden
  - link anomaliebeheer: slechts indien definities gevalideerd → formele anomalie, detectie mogelijk

Data Matching algorithms

- Matching on two levels
  - Field-per-field
  - Per record (aggregation)
- Performance [!]
  - Compairing N records with each other
  - N(N-1)/2



WINKLER W. E., *Overview of Record Linkage and Current Research Directions*, 2006. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

Data Matching algorithms
Field-per-field

- Character strings
  (names, streetnames, numbers, … wherever there is fuzziness)

- Two fields are “the same”
  - AFSCA "=" Agence Fédérale pour la Sécurité de la Chaîne Alimentaire
  - Yves "=" Yves
  - Yves "=" yves
  - Yves "=" Yvse
  - Bontemps "=" Bontan
  - Bontemps, Yves "=" Yves Bontemps

Data Matching algorithms – Field per field

Classes of "comparison routines"

	Rules & predicates	Phonetics
Booleans (classifiers)	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*
	Word-based	**Token-based**
Similarity-based	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosine • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching algorithms – Field per field

Classes of "comparison routines" – Booleans: Rules & predicates

- Boolean relations
  - is-equal-to
  - is-a-prefix-of
  - is-a-suffix-of
  - is-an-abbreviation-of
  - stems-from
  - …

- Exemples:

  Bontemps ~ Bontemps

  Bont ~ Bontemps

  Emps ~ Bontemps

  S.A. ~ Soc. Anon.

  Bakkerij ~ Bakker

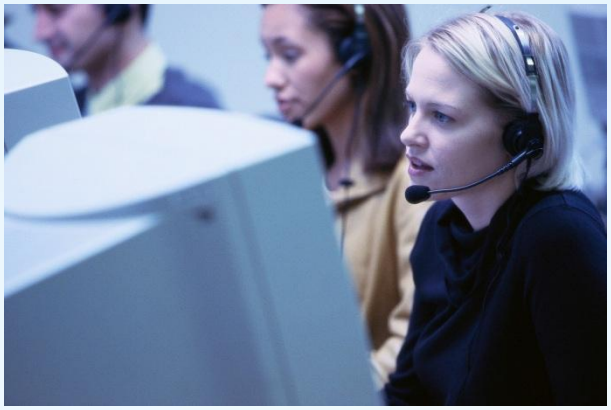
Data Matching algorithms – Field per field

Classes of "comparison routines" – Booleans: Rules & Predicates

- Hand-coded rules
  - Hard-wired (Java, C, COBOL, …)
  - Rule engine (declarative rules: pre-defined, user-defined)
- May be packaged in commercial software (black-box?)
- ☺ Making the best use of domain knowledge is possible
- ☹ Fastidious
- ☹ Costly (in development & in maintenance)

Data Matching algorithms – Field per field

Classes of "comparison routines" – Booleans: Phonetics



- Origin
  - transcription errors: oral → written
- Ex: Bontemps:
  - Montand
  - Bontant
  - Bonton
  - Bontand
  - Bautemps
  - …

- Principle:
  - word m → representation of its pronunciation $p(m)$
  - $p(m) = p(m') \Leftrightarrow m \sim m'$

Data Matching algorithms – Field per field

Classes of "comparison routines" – Booleans: Phonetics

- Russel Soundex Algorithm (1918)
  - Garder première lettre
  - Supprimer a,e,h,i,o,u,w,y
  - Codage:
    1: B,F,P,V
    2: C,G,J,K,Q,S,X
    3: D,T
    4: L
    5: M, N
    6: R
  - Garder 4 premiers symboles (padding avec 0)

- Exemple
  - Bontemps
    BNTMPS
    B53512
    B535
  - Bondant
    BNDNT
    B5353
    B535

Data Matching algorithms – Field per field

Classes of "comparison routines" – Booleans: Phonetics

- Alternatives:
  - Metaphone et Double Metaphone
  - NYSIIS et NameSearch®
  - Daitch-Mokotoff
    - Slavic & German languages
    - 54 entries
  - Fonem
    - French-oriented
    - 64 rules
  - Phonex, etc…

Data Matching algorithms – Field per field

Classes of "comparison routines"

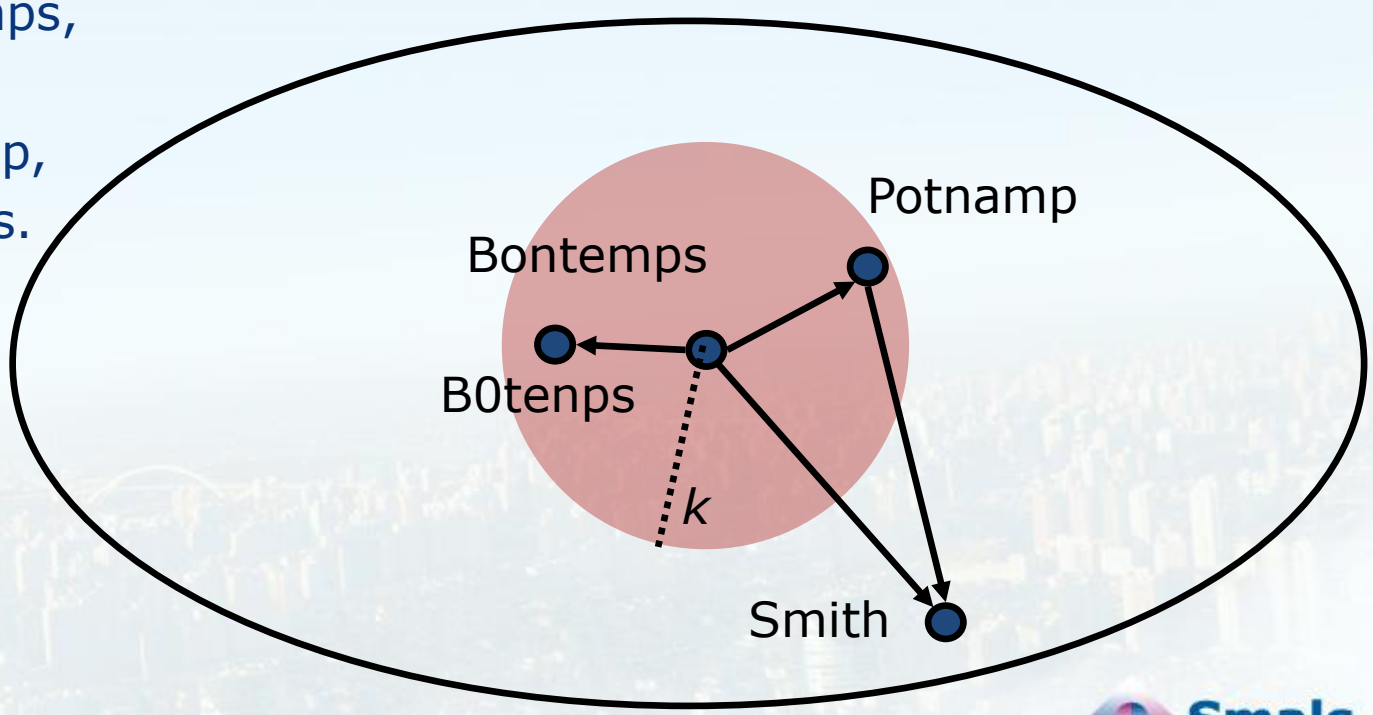
	Rules & predicates	Phonetics
Booleans (classifiers)	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*
Similarity-based	**Word-based** • Edit distances • Lettres communes *Bontemps ~ Bnotmps*	**Token-based** • Cosine • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity-based

- Remplacer l'approche booléenne (0,1) par une approche plus fine

- Idée: Similarité variable entre
  - Bontemps,
  - Smith,
  - Potnamp,
  - B0tenps.



Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity: Word-based

Edit Distance

- Distance between two strings = number of « operations » to transform one into the other
  - Insertion (I)
  - Deletion (D)
  - Substitution (S)
- Cost per "error" (operation)
  - OCR, Clavier, etc.
- Most popular: Levenshtein
  - http://en.wikipedia.org/wiki/Levenshtein_distance
- Exemple
  - Bontemps
  - Bnontemps (I)
  - Bnontemps (D)
  - Bnontemps (D)

➔ 3 operations

Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity: Word-based

Jaro-Winkler Distance

- Longest string counts *s*
  - window of *(s/2) – 1*
- Exemple
  - *s* = 7 → window = 3
  - *m* = 2 characters are in matching positions
  - *t* = 1 further "transposition" within window
  - *d = 1/3 * (m/s1 + m/s2 + (m-t)/m)*
  - d = 1/3 * (2/7 + 2/7 + (2–1)/2 = 0.221

	S	L	U	I	T	E	N
S	1						
T					1		
R							
U			1				
D							
E						1	
L							

REF: http://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity: Token-based

- Token-based metrics
  - When?
    - Multiple elements in compared fields
  - Exemples
    - Ignoring word order
      - {Bontemps, Yves} = {Yves, Bontemps}
    - Jaccard metric

$$\frac{|S \cap T|}{|S \cup T|}$$

      - {{Bontemps;Yves}} {{Yves; Bontemps}}
        → 2/2 = 1

Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity: Token-based

- Often: weighting schemes are used
  - E.g. Term Frequency / Inverse Document Frequency (TF-IDF)
  - Uses Rarity of a word as a measure for importance

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

- Exemples:
  - Organisme National des Déchets Radioactifs et des Matières Fissiles
  - Office Nationale des Matières Fissiles et Déchets Radioactifs
  - Office Nationale de la Sécurité Sociale

Data Matching algorithms – Field per field

Classes of "comparison routines" – Similarity: Token-based

- Recursive approaches:
  - Use "word-based" routine (Levenshtein, …) to determine similarity between words (*Office* ~ *Ofice*)
  - Use "token-based" technique to combine the results
  - Ref: Soft TF-IDF, etc.

Data Matching
En pratique

- Comparaison phonétique (KBO):
  - Décomposition en mots
  - Mots mis sous forme phonétique.
    - Variante de Soundex
    - Ignore termes habituels (S.A., ...)
    - Stemming
  - Comparaison
    (prise en compte de l'ordre des mots)

Data Matching algorithms

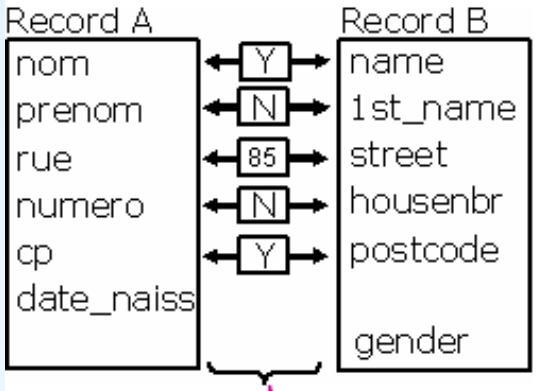
Aggregate level – "Per record"

- Approaches:
  - Deterministic
    - It is always clear why the algorithms decided → match or non-match
  - Probabilistic (*global scoring*)
    - Interpretation is difficult, if not impossible
    - Fellegi & Sunter, 1969
      - Cfr. Machine Learning
      - Pour chaque comparaison, deux probabilités:
        - m = Prob. de Y si A et B sont un vrai match
        - u = Prob. de Y si A et B sont un vrai unmatch
      - Le score est calculé sur base de ces probabilités (m/u)
      - Classification en fonction du score

Data Matching algorithms

Aggregate level – "Per record"

résultat champ par champ



résultat comparaison par record

Approche déterministe

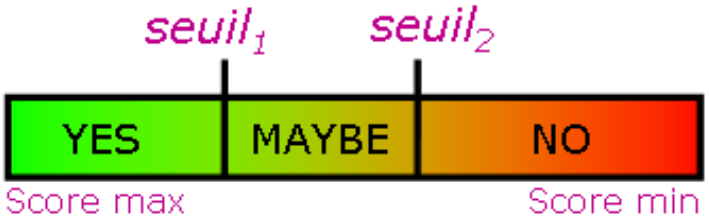
	Nom	Prénom	Rue	Num	Postcd	Resultat
1	Y	Y	-	-	Y	Match
2	N	-	-	-	-	No
3	Y	N	90	Y	Y	Match
4	Y	N	80	N	Y	Maybe
5	-	-	-	-	-	No

Approche probabiliste

- *Poids*
  - Nom = 0.95
  - Prenom = 0.45
  - Rue = 0.25
  - Num = 0.01
  - CP = 0.05

  ➔ *Score = 1.25*

- *Règle :*



Data Matching - Performance

- Matching = comparing all records of DB A with all records of DB B
  (or comparing all records of a DB with each other)
- E.g.
  - A : 100 000 rec.
  - B : 100 000 rec.

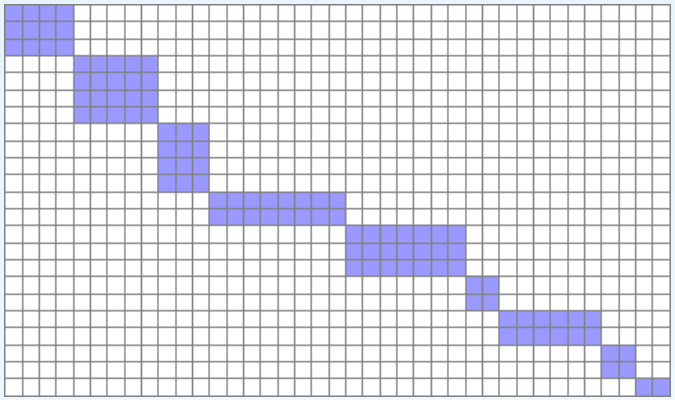
  → 10 000 000 000 comparisons to calculate !
- De-duplication or double-detection:
  - N records

  → N(N-1)/2 comparisons to calculate !

➔ Optimisations nécessaires…

Data Matching - Performance

- **Blocking** technique:
  - A critical dimension is chosen → to form blocks or windows
  - Records are only compared within windows
  - Ex: Code Postal, …
  - Trade-off between precision and performance
  - Therefore, the dimension chosen must be highly trustworthy (*of good data quality*)
  - Requires indexation and sorting
  - Multi-passes, or the combination of multiple windowing techniques, are possible



Data Matching – Performance
Illustration "Window Key" creation

Window Key Rule for Window Key Generator Window Key 01 - Rule 1

Primary Attribute	TQ_GOUT_POSTAL_CODE
No. of characters	2
Selection	Numerics only
Secondary Attribute	TQ_GOUT_POSTAL_CITY
Condition	Blanks or Zeros
No. of characters	2
Selection	Numerics only

Back

Window Key Rule for Window Key Generator Window Key 01 - Rule 2

Primary Attribute	TQ_GOUT_STREET_NAME
No. of characters	2
Selection	First character + subsequent non-repeating consonants
Secondary Attribute	PR_BOX1_NUMBER
Condition	Blanks
No. of characters	1
Selection	Numerics only

Back

Window Key Rule for Window Key Generator Window Key 02 - Rule 1

Primary Attribute	PR_BUSNAME_RECODED_01
No. of characters	3
Selection	First character + subsequent non-repeating consonants
Secondary Attribute	
Condition	
No. of characters	
Selection	

Back

Selection

- Any character
- Any character (from the last word)
- Vowels only
- Vowels only (from the last word)
- Non-repeating vowels only
- Non-repeating vowels only (from the last word)
- Consonants only
- Consonants only (from the last word)
- Non-repeating consonants only
- Non-repeating consonants only (from the last word)
- First character + subsequent vowels
- First character + subsequent vowels (from the last word)
- First character + subsequent non-repeating vowels
- First character + subsequent non-repeating vowels (from the last word)
- First character + subsequent consonants after hyphen
- First character + subsequent consonants
- First character + subsequent consonants (from the last word)

Data matching

Real world example: confrontation of 2 DB

Source	match type	Denomination	Adres	Boite	Postcd	Commune	Cdpays
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
R	115	PROJEKT SERWIS (LUTY WANDA)	UL. BOHATEROW MODLINA 63	42	05-100	NOWY DROW MZAOWIE	PL
R	115	PROJEKT SERWIS LUTY WANDA NOWY DWOR	UL BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZOWIE	PL
R	135	PROJEKT SERWIS LUTY WANDA	BOHATEROW MODLINA 63/43		05-100	NOWY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATERROW MODLINA 63/43	N/A	05-100	NOWY DWOR MAZOWIE	PL

2 databanken, L en R

- ✓ er bestaat geen 'vreemde sleutel'-relatie tussen L en R
- ✓ DQTool detecteert dubbels en organiseert in clusters
- ✓ DQTool legt link tussen beide databanken
- ✓ mogelijke fraude ?

Data matching

Real world example – importance of internal standardisation step

C Postcode	Tq Gout Postal Code	Straatnaam Voll	Tq Gout Street Name	Huisnummer	Pr House N...	Gemeentenaam	Tq Gout Postal City
1020	1020	RUE E VANDER AA	RUE ERNEST VANDER AA	1	1	Brussel	BRUSSEL
1020	1020	rue Vander Aa	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	91 R VAN AA	RUE VAN AA		91	Elsene	ELSENE
1050	1050	27 R.VAN AA	RUE VAN AA		27	Elsene	ELSENE
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Ixelles	IXELLES
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Elsene	ELSENE
1020	1050	rue Van Aa	RUE VAN AA	2	2	Bruxelles	IXELLES
1050	1050	2 R VAN AA	RUE VAN AA		2	Ixelles	BRUXELLES
1000	1000	R JOSEPH II 40	RUE JOSEPH II		40	Bruxelles	BRUXELLES
1000	1000	rue Joseph II 71 (...	RUE JOSEPH II		71	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
1040	1000	Rue Joseph II 5-7	RUE JOSEPH II		5-7	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II 67A	RUE JOSEPH II		67A	Bruxelles	BRUXELLES
1030	1000	rue JOSEPH II, 114 -	RUE JOSEPH II	116	114 - 116	Schaarbeek	BRUXELLES

Resultaat (adresvalidatie, adrescleansing)

- ✓ gecorrigeerde postcode
- ✓ gestandaardiseerde straatnaam
- ✓ correct ingedeelde adreselementen (parsing)
- ✓ gecorrigeerde gemeentenaam
- ✓ dubbels gedetecteerd en georganiseerd in clusters

Data standardisation & matching

Not only names & addresses …

P/N	DESCRIPTION
1774-5674	TUBE, CENTRIFUGE POLY S 15ML (CS/500)CONICAL-BOTTOM
1774-5675	TUBE, CENTRIFUGE PPL 15ML (CS/500)CONICAL-BOTTOM
1774-4532	TUBE, CENTRIFUGE PPL 50ML (CS/500)CONICAL-BTTMPCK 25/RACK
1774-4538	TUBE, CENTRIFUGE POLY S 50ML (CS/500)CONICAL-BTMPK 25/RACK
645-4556	PIPET, CLEAR SEROLOGICAL 2ML (CASE/500)
195-7934	NUT, LOCK RH,11"
3324-7955	VIAL, WHEATON 33* CLEAR 4ML (CS/144)

P/N	ITEM NAME	MATERIAL	SIZE	UOM	DESCRIPTOR	PACKAGE	PACK METHOD
1774-5674	CENTRIFUGE TUBE	POLYSTERENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-5675	CENTRIFUGE TUBE	POLYPROPILENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-4532	CENTRIFUGE TUBE	POLYPROPILENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
1774-4538	CENTRIFUGE TUBE	POLYSTERENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
0645-4556	SEROLOGICAL PIPET		2	ML	CLEAR	CASE/500	
0195-7934	LOCK NUT		11	IN	RIGHT HAND		
3324-7955	WHEATON VIAL		4	ML	CLEAR	CASE/144	

Data Quality Tools supporting iterative concertation with the business – real Address Cleansing project example

Match P.	Fst Denom Lnm	Street Lnm	Street Num	City Lnm	Postcode
	inter connector zeebrugge terminal sc/cv	8th floor, aldwych	61	london	0000
402	interconnector zeebrugge terminal sc/cv	8th floor aldwych	61	wc24ae london	0000
	HAAN TECHNIEK	SAAL VAN ZWANE...	2	TILBURG	0000
402	Haan techniek □/n	Saal Van Zwanenbe...	2	Tilburg	0000
	Brixx Sp. z.o.o.	Ul. Algierska	8	Warszawa	03-977
402	BRIXX sp.z.0.0.	Ul. Algierska	8	Warszawa	03*977
110	Brixx Sp.z.o.o.	Ul Algierska	8	Warszawa	03-977
110	Brixx sp z.o.o.	Ul. Algierska	8	Warszawa	03-977
135	BRIXX SP.Z.O.O. □/nond.nr. PL - 113-2...	ul. ALGIERSKA	8	WARSZAWA	03-977
110	Brixx Sp.z.oo	ul.Algierska	8	Warszawa	03977
135	BRIXX SP ZOO □/nPL 1132760927	UL ALGIERSKA	8	WARSZAWA	03-977

Project | Data rows filtered by selected values in 'Highest Lev1matchpat' | rows: 26662

- Elk match pattern is een hulp voor de analist bij het finetunen van matching routines (Discovery)
- Iteratief overleg met business i.v.m. definitie en afhandeling van elk match pattern (Validation) → correctiescenario's
  - **110**: betrouwbaar
  - **402**: afwijking in de benaming + probleem met de postcode
  - **135**: grotere afwijking in de benaming
- Elk gevalideerd match pattern → anomalie (anomaly_type_id)

Smals ICT for society

DQ tools support data quality problem resolution
e.g. Integrating incoherent sources

Bv.: 'commonize' telefoonnummer voor alle matched records

Date	First	Last	Phone	Email	Source
08/02/00	Art	Barrios	908-845-1234	bigwheels@hotmail.com	WEB
12/02/2005	A.	Barros	908-845-1234	abarrios@accen.com	CRM
6/17/2003	Arthur	Barrios	~~(902)-845-4417~~	abarrios@accen.com	SAP

Specific matching routines

Date	First Name	Last Name	Ignore Punctuation	Absolute

Distinct survivorship routines

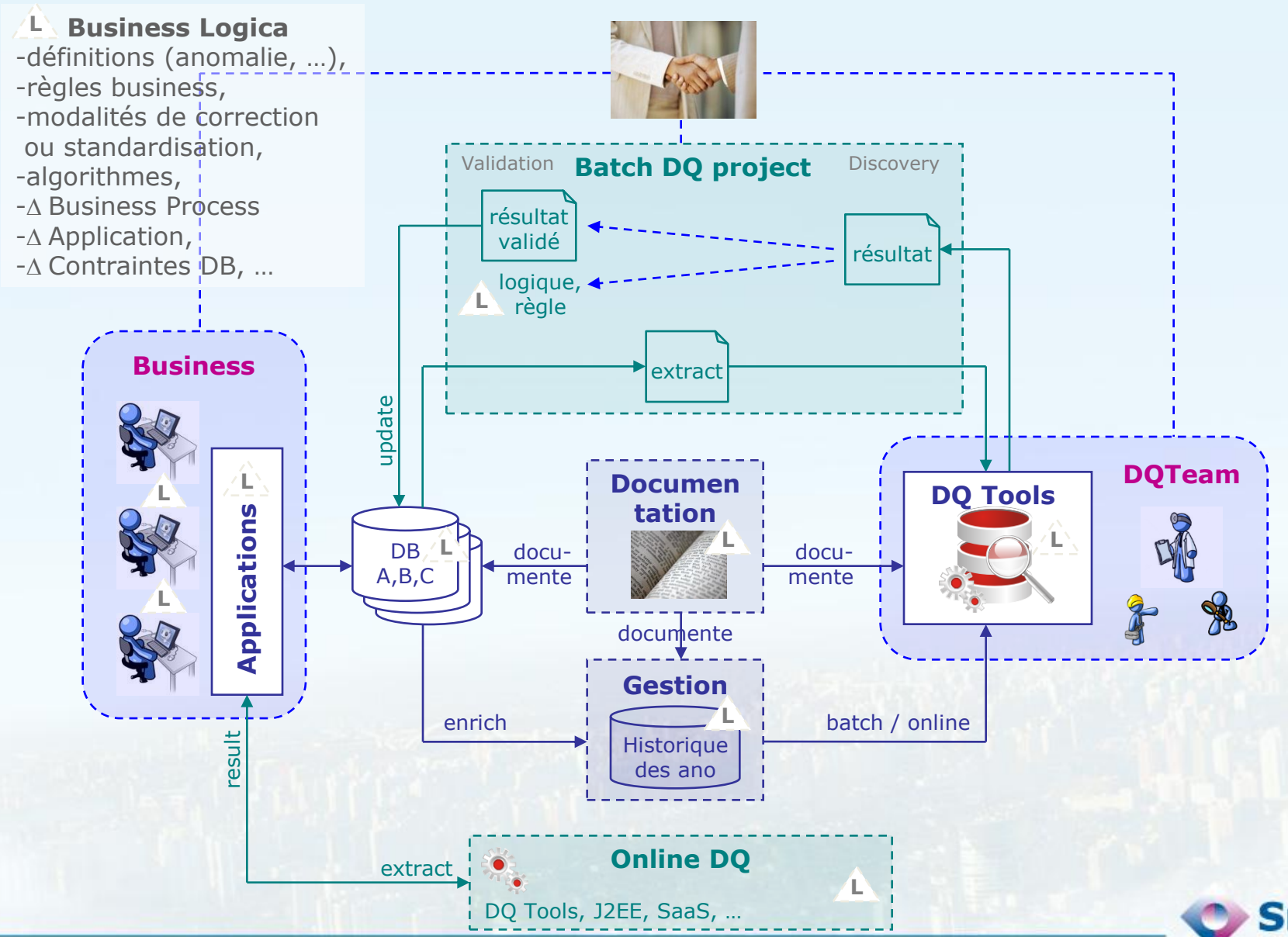
Most Recent	Complete	Most Common	Most Recent	Best Source

12/2/2005	Arthur	Barrios	908-845-1234	abarrios@accen.com

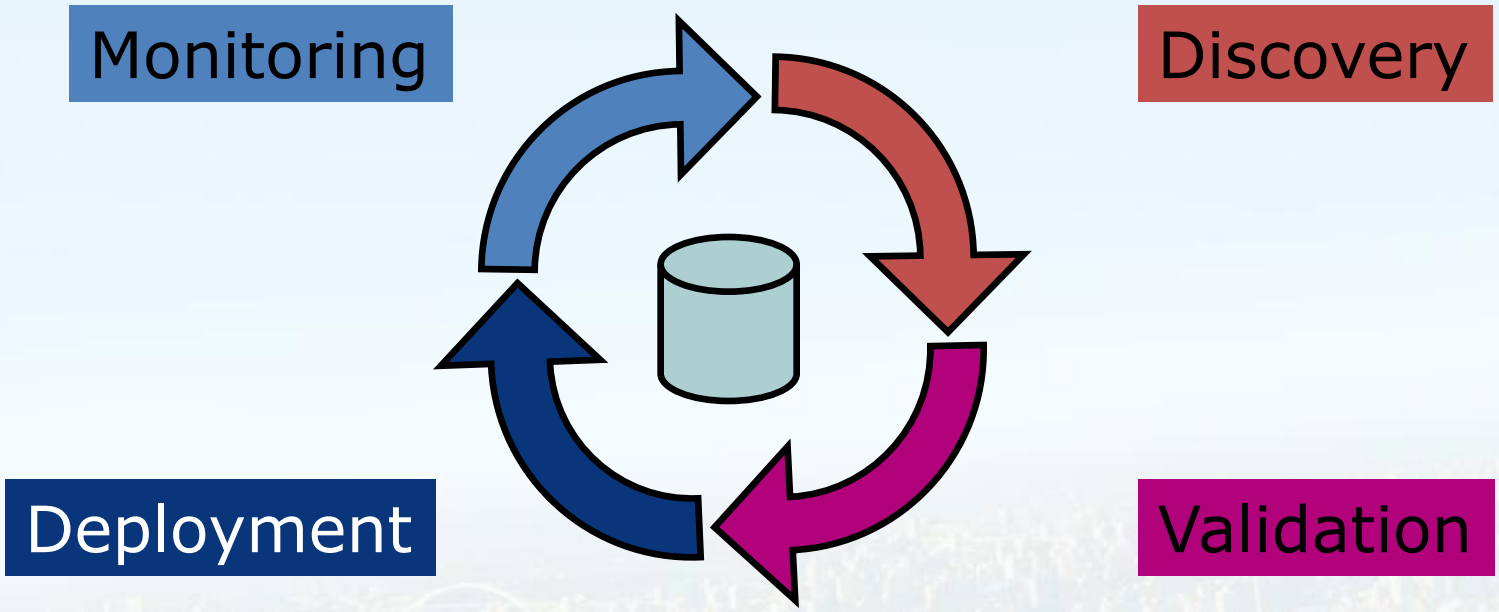
Contents

- Introduction – Smals
- Introduction – The Reality of Data Quality
- Functionalities of Data Quality Tools
- Organization for Data Governance
- Conclusions

Organization for Data Governance – Smals project approach



Data Quality Improvement - continuous cycle



Organization for Data Governance

- Data Governance Projects: phases
  - Discovery
  - Validation (in close concert with the business)
  - Deployment (what, how (solution architecture))
  - Governance (monitoring, lifecycle (incl. retirement))

Contents

- Introduction – Smals
- Introduction – The Reality of Data Quality
- Functionalities of Data Quality Tools
- Organization for Data Governance
- Conclusions

Conclusions – When to act (on DQ)

- Whenever the quality of the data has strategic importance
- Whenever the (low) quality of data entails recurrent costs and limits business capabilities

consider
fitness for use

- Whenever the *use* of the data *changes*
  - Typical context: projects & governance progams, involving
    - Data Integration, Data Migration, Reengineering
    - Fuzzy matching (e.g. names, adresses)
    - Standardisation, Master Data Management
    - Incoherence detection (even fraud) & resolution
    - Anomaly management

Conclusions – How to act

- Because of typical DQ project nature
  - three 'phases'
    - Discovery
    - Validation
    - Deployment

Iterative process with business involvement ← critical success factor

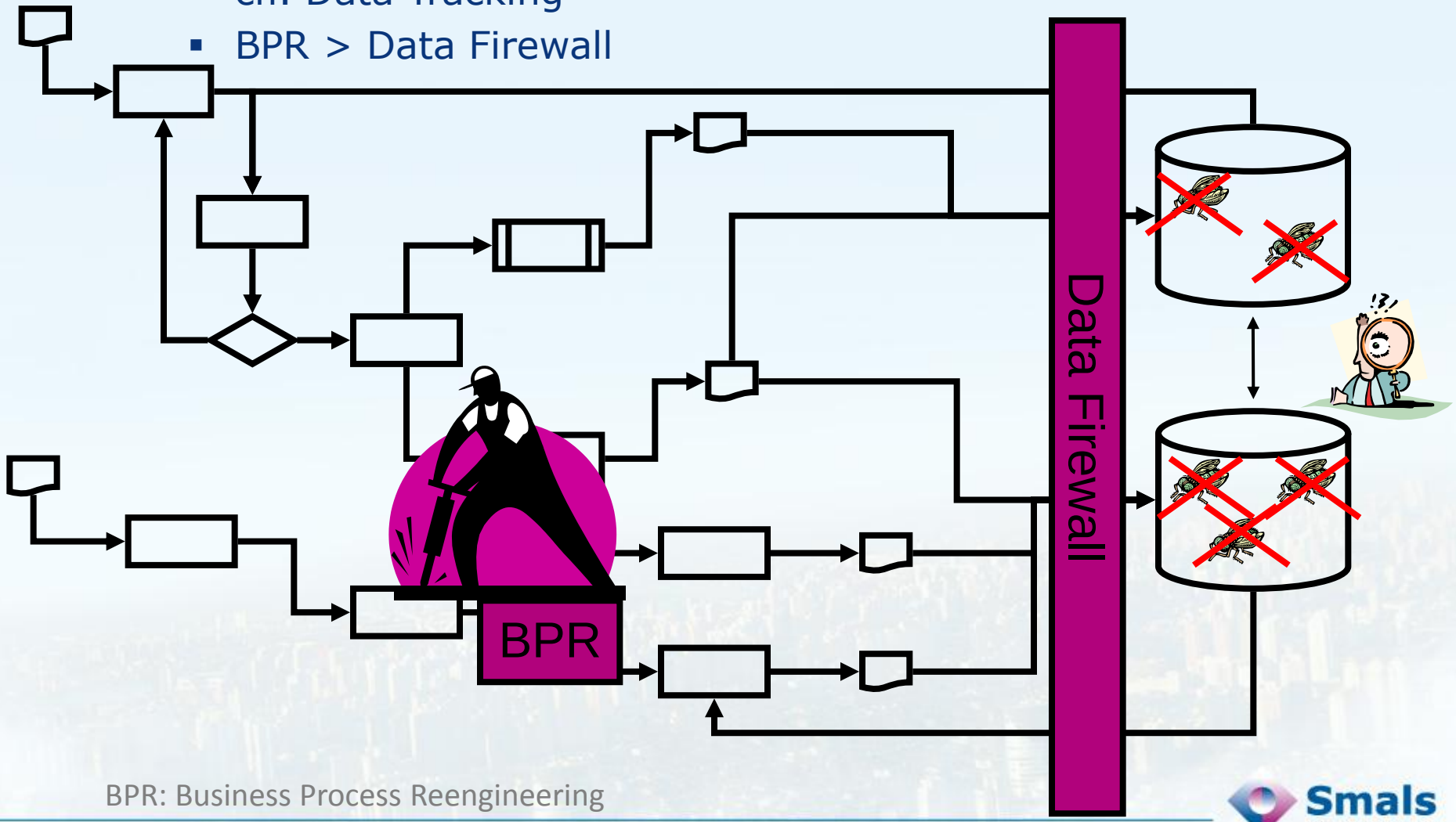
- Data Quality Tools are better at this, thanks to:
  - Performantie
  - Productiviteit
  - Flexibiliteit

- configure instead of developing code
- data matching >= 1 million records/hr
- GUI/visualize/drill-down/export
  - format = ideal for business concertation

In our experience, Data quality tools enable users to efficiently tackle a wide range of data quality problems. Without such tools, it would be costly, time consuming, or even impossible to remediate such problems.

Conclusions – Where to act

- Act at the source (of problems)
  - cfr. Data Tracking
  - BPR > Data Firewall



BPR: Business Process Reengineering

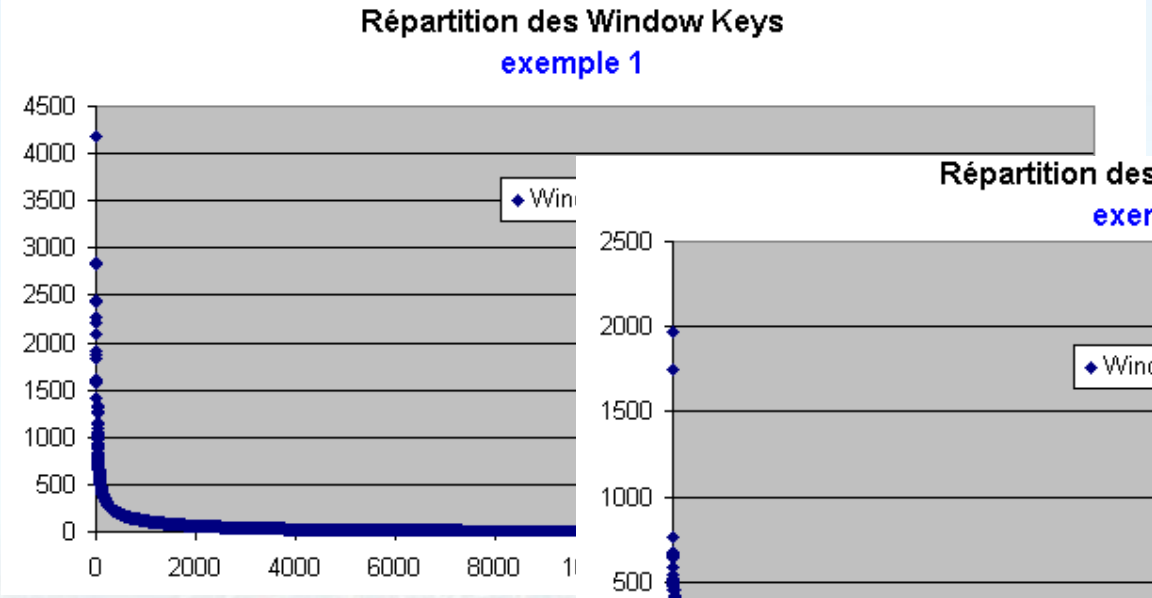
Annex

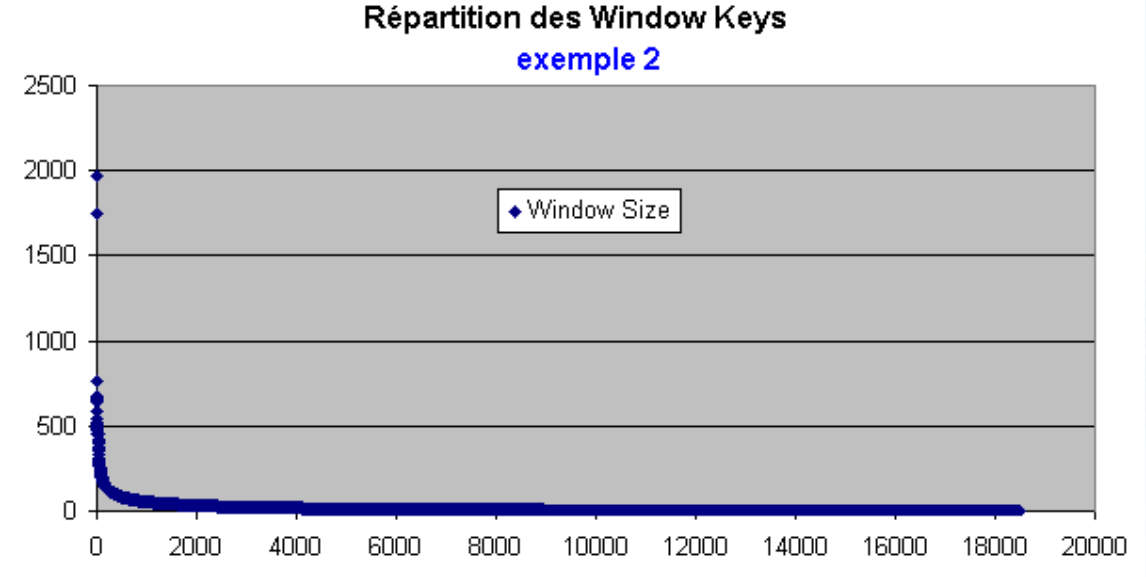
- Window Key creation
  - Character choice routines

Any character
Any character (from the last word)
Vowels only
Vowels only (from the last word)
Non-repeating vowels only
Non-repeating vowels only (from the last word)
Consonants only
Consonants only (from the last word)
Non-repeating consonants only
Non-repeating consonants only (from the last word)
First character + subsequent vowels
First character + subsequent vowels (from the last word)
First character + subsequent non-repeating vowels
First character + subsequent non-repeating vowels (from the last word)
First character + subsequent consonants after hyphen
First character + subsequent consonants
First character + subsequent consonants (from the last word)
First character + subsequent non-repeating consonants
First character + subsequent non-repeating consonants (from the last word)
Initials
Initials (sorted)
Initials (from the last word)
If input value is 1, store value as 1. All other input values recode to a 2.
Numerics only
Non-repeating numerics only
Algorithm based on standard NYSIIS code
Original soundex
Reversed original soundex
Reversed improved soundex
Improved soundex
Any sorted
Vowels only sorted
Non-repeating vowels only sorted
Consonants only sorted
Non-repeating consonants only sorted
First characters + subsequent vowels sorted
First characters + subsequent non-repeating vowels sorted
First characters + subsequent consonants sorted
First characters + subsequent non-repeating consonants sorted
Initials sorted
Blanks
Zeros
Blanks or Zeros
Reversed NYSIIS code

Annex

- Window Key Size distribution is important
  - Matching complexity = o([window size]²)





Annex

- Field by field comparison routines
  - Commercial tool example (deterministic)

Comparison Routines

Routine	Purpose
ABSOLUTE	Exact match comparison and score determination.
APTNO	Apartment number field comparison and score determination.
ARRAY1	Compare segments of a field to segments of another field, and determine the relationship to blank values.
ARRAY2	Compare segments of a field to segments of another field, and determine the relationship to blank values.
BUSNAME	Business name field comparison and score determination.
CHARACTER	Compares two character fields for Asian countries.
DATE	Comparison of two date fields.
DIFFER	Numeric field comparison and score determination.
FLAG10	On/off(0/1) field comparison and score determination.
FLAGFM	On/off(F/M) field comparison and score determination for gender.
FLAGGN	On/off(G/N) field comparison and score determination for gender.
FLAGMF	On/off(M/F) field comparison and score determination for gender.
FLAGYN	On/off(Y/N) field comparison and score determination for gender.
FRSTNAME	First name field comparison and score determination.
GENER	Generation field comparison and score determination.
HOUSENO	House number field comparison and score determination.
MXDNAME	Substring presence in fields of mixed name form records.
NYSIIS	Matches two strings of data, using an algorithm based on a standard NYSIIS algorithm.
RNYSIIS	Links two strings using an algorithm based on a standard NYSIIS algorithm.
ONECOM	Test for the Presence of one commercial record.
PARTIAL1	Blank field comparison and score determination.
PARTIAL2	Blank field comparison and score determination.
POSTCODE	Postal code field comparison and score determination.
PREFIX	Prefix field comparison and score determination.
PREVENT	Prevents matching on fields named in Field/Comparison Routine Listing.
SOCSEC	Social security field comparison and score determination.
SOUNDEX1	Letter comparison and score determination.
SOUNDEX2	Letter comparison and score determination using an improved SOUNDEX algorithm.
SPELLING	General purpose text comparison and score determination.
STATUS	Blank field comparison and score determination.
STREETS	Street name field comparison and score determination.
SUBSTRNG	Substring presence and score determination.
TWORET	Test for the presence of two retail records.

OK Cancel

Annex

- Dimensions of Data Quality
  - *Relevant*
  - *Accurate*
  - *Timely*
  - *Complete*
  - *Understood*
  - *Trustworthy*