Data Quality Tools

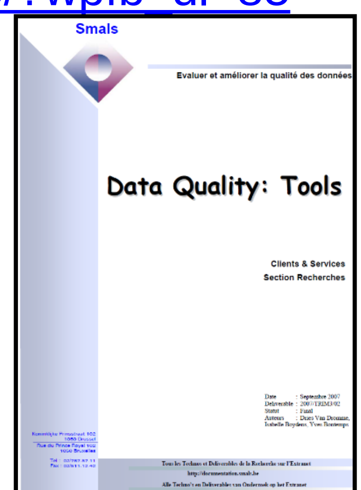
Concepts and practical lessons from a vast operational environment

Gani Hamiti – 13/03/2018 @ ULB

Before we start…

Reference documents – Smals Research (Isabelle Boydens, Yves Bontemps, Dries Van Dromme) about data quality & DQ tools

- Gestion intégrée des anomalies
  - https://www.smalsresearch.be/?wpfb_dl=62
- Data quality tools :
  - https://www.smalsresearch.be/?wpfb_dl=85



Before we start…

- Although technical matter, hand in hand with application area specialists (« business » in the uncommercial sense)

- Each time iterations with application area specialists are crucial, logo on **upper right corner**:

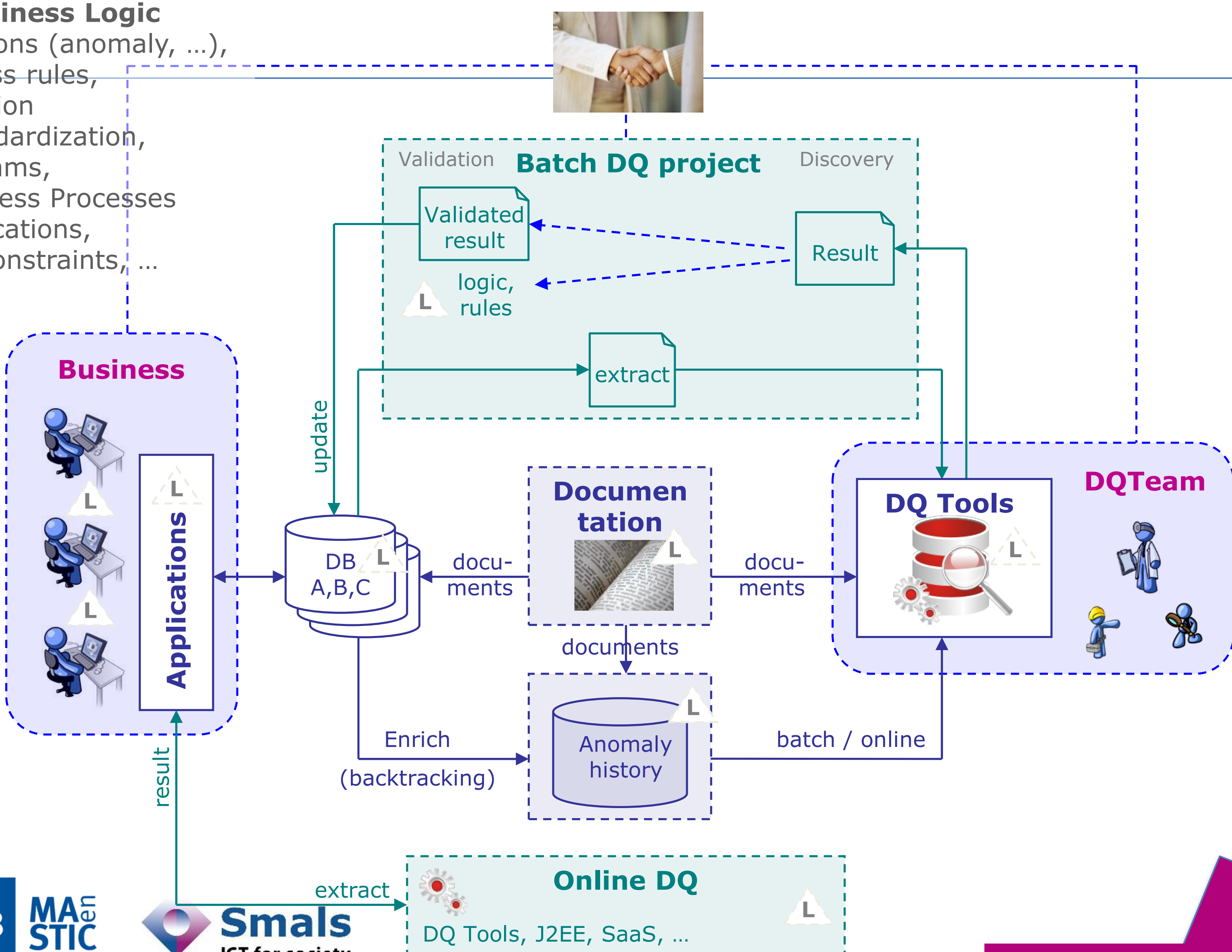
Contents

Contents

Contents

1. Preventive and curative approaches : organization

Business Logic
- -definitions (anomaly, …),
- -business rules,
- -correction
  or standardization,
- -algorithms,
- -Δ Business Processes
- -Δ Applications,
- -Δ DB constraints, …



Source: Y. Bontemps, D. Van Dromme

Contents

2. The curative approach

- Profiling: what’s happening into our data and metadata (if available) ?
  - Investigate DQ and analyze (un)known anomalies
  - Measure when possible

2. The curative approach

- Profiling: what's happening into our data and metadata (if available) ?
- Standardization: build and apply standards to our data
  - Formal or fundamental
  - Enriching with Knowledge DBs and/or Business Rules

2. The curative approach

- Profiling: what’s happening into our data and metadata (if available) ?
- Standardization: build and apply standards to our data
- Matching: which records belong together?
  - Detect duplicates and inconsistencies: variable fuzziness
  - Deduplicate
  - Chose or build a « golden record »
  - /!\ Performance

2. The curative approach

- Profiling: what's happening into our data and metadata (if available) ?
- Standardization: build and apply standards to our data
- Matching: which records belong together?

2. The curative approach

- Profiling: what’s happening into our data and metadata (if available) ?
- Standardization: build and apply standards to our data
- Matching: which records belong together?
- Dedicated tools, specific to one area or « all-in-one »

2. The curative approach: DQ tools

2. The curative approach: DQ tools

- Since 1980’s, initial core business: names and addresses
  - Ever-present issue
  - Ubiquitous: companies, client data, service providers, B2B, public administrations…

2. The curative approach: DQ tools

- Since 1980's, initial core business: names and addresses
  - Ever-present issue
  - Ubiquitous: companies, client data, service providers, B2B, public administrations…
- Complex and changing standards
  - Knowledge bases built over time
  - Taking international context into account
  - Regular updates

2. The curative approach: DQ tools

- Since 1980’s, initial core business: names and addresses
  - Ever-present issue
  - Ubiquitous: companies, client data, service providers, B2B, public administrations…
- Complex and changing standards
  - Knowledge bases built over time
  - Taking international context into account
  - Regular updates
- Today, extended to all alphanum strings
  - Thousands of mature algorithms
  - Decades of optimizations

2. The curative approach: DQ tools

- Since 1980's, initial core business: names and addresses
  - Ever-present issue
  - Ubiquitous: companies, client data, service providers, B2B, public administrations…
- Complex and changing standards
  - Knowledge bases built over time
  - Taking international context into account
  - Regular updates
- Today, extended to all alphanum strings
  - Thousands of mature algorithms
  - Decades of optimizations
- Adapted to DQ work nature
  - Iterations and drill-down
  - Constant business involvement (critical!)
  - Less time wasted in development: more efficient resource distribution

Contents

3. DQ@Smals: projects (1)

- **2010**
  - **FOD Economie**: KBO Adreskwaliteit
  - **RSZ**: 30bis werfmeldingen – dubbeldetectie
  - **eHealth-platform**: inconsistence management (multiple DBs)
  - **FAGG**: Datamigratie Kadaster Officina
- **2011**
  - **FAGG**: Datamigratie Kadaster Officina
  - **SIGeDIS**: 2de pensioenpijler - preload KBO
  - **VAZG**: Datakwaliteit Vaccinnet
- **2012**
  - **eHealthPlatform** – opbouw van Validated Authentic Souces (VAS)
  - **RSZ** sociale-fraudebestrijding
- **2013**
  - **eHealthPlatform** – VAS (continued)
  - **RSZ** sociale-fraudebestrijding (adresmatching)
  - **RSZ** fuzzy matching Limosa-kadaster (foreign employees in Belgium)

3. DQ@Smals: projects (2)

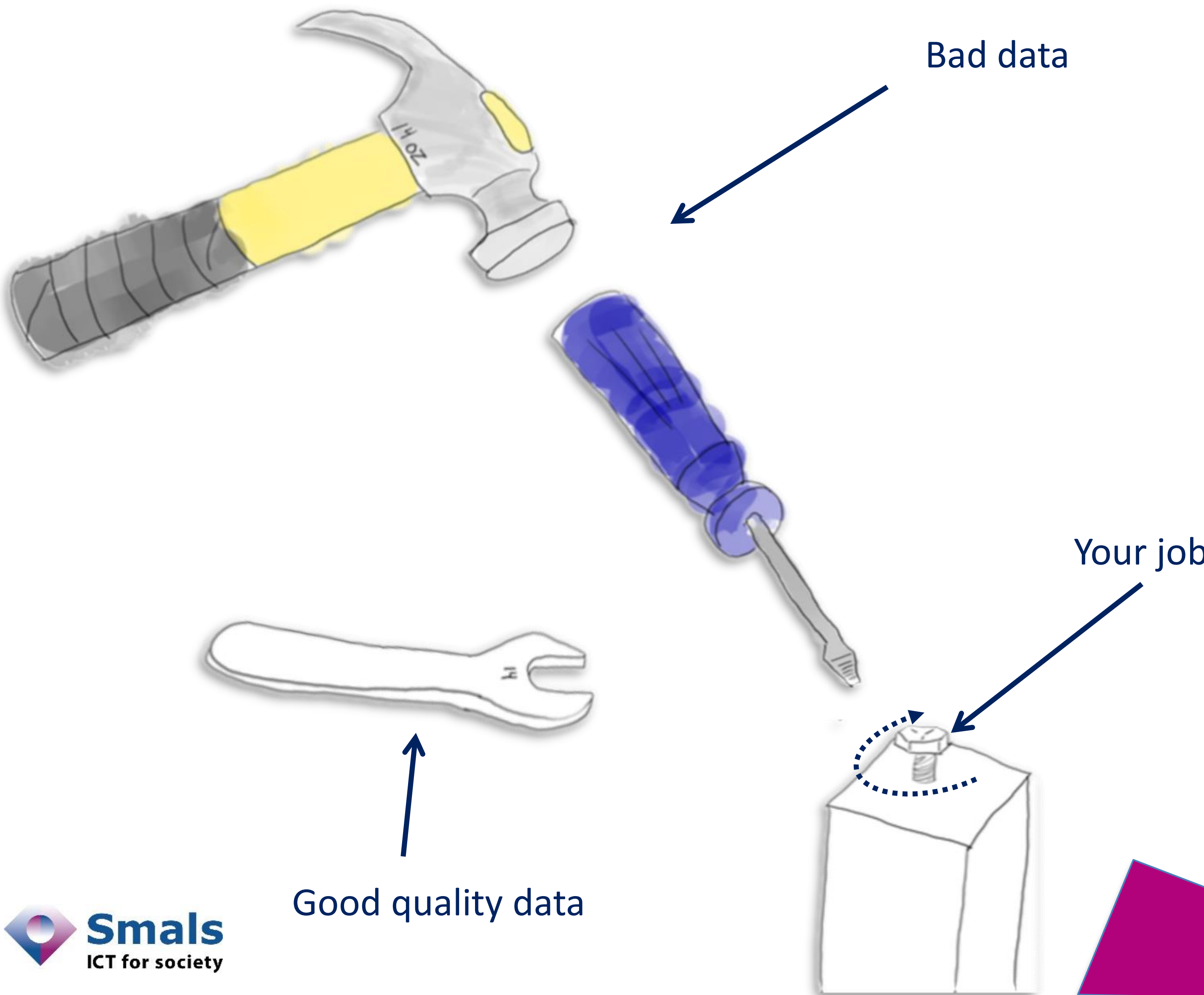
- **2014-2015-2016**
  - 2015-2016: **RSZ** EDE (Dossier Electronique de l'Employeur)
  - **eHP** – VAS (continued)
  - **RSZ** sociale-fraudebestrijding : matching entities from various authentic sources (continued)
- **2017-2019**
  - **KBO – Repertorium** : comparative profiling
  - **RSZ** – Directie Risicobeheer : register matching enterprises from various authentic sources
  - **eHP** – VAS (continued)
  - **Fédération Wallonie-Bruxelles:** data quality management and integration from various financial and accounting databases
  - **Migration OSSOM – ONSS :** inconsistency detection and data migration
  - **FoLeEn**: repertory to identify foreign companies

Contents

ULB MA en STIC

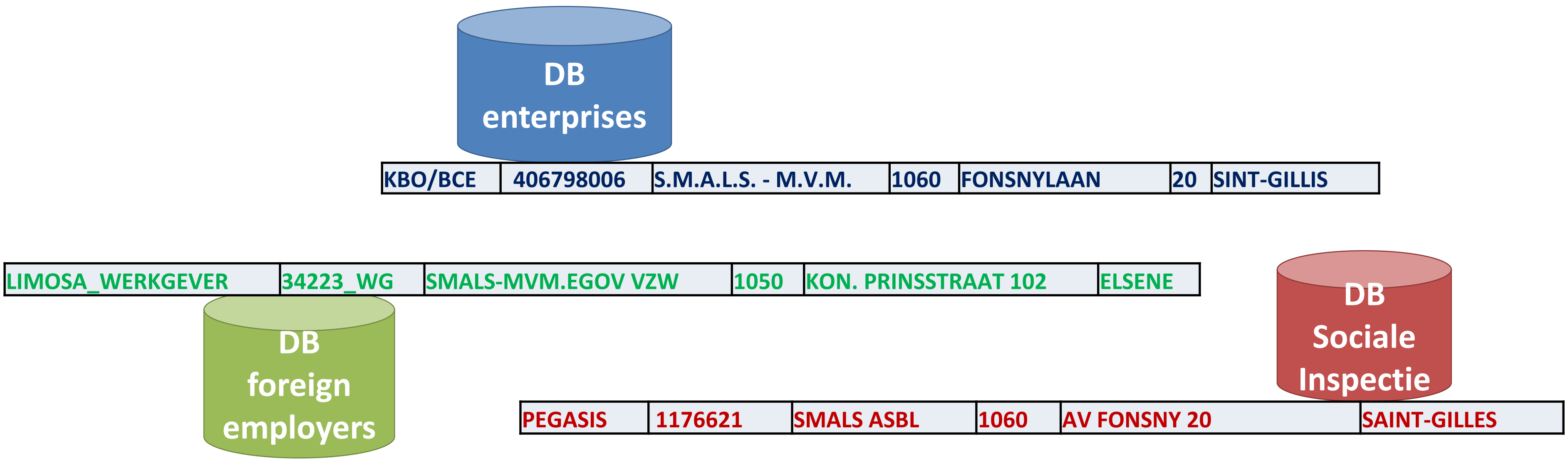
Smals ICT for society

4. Fitness for use



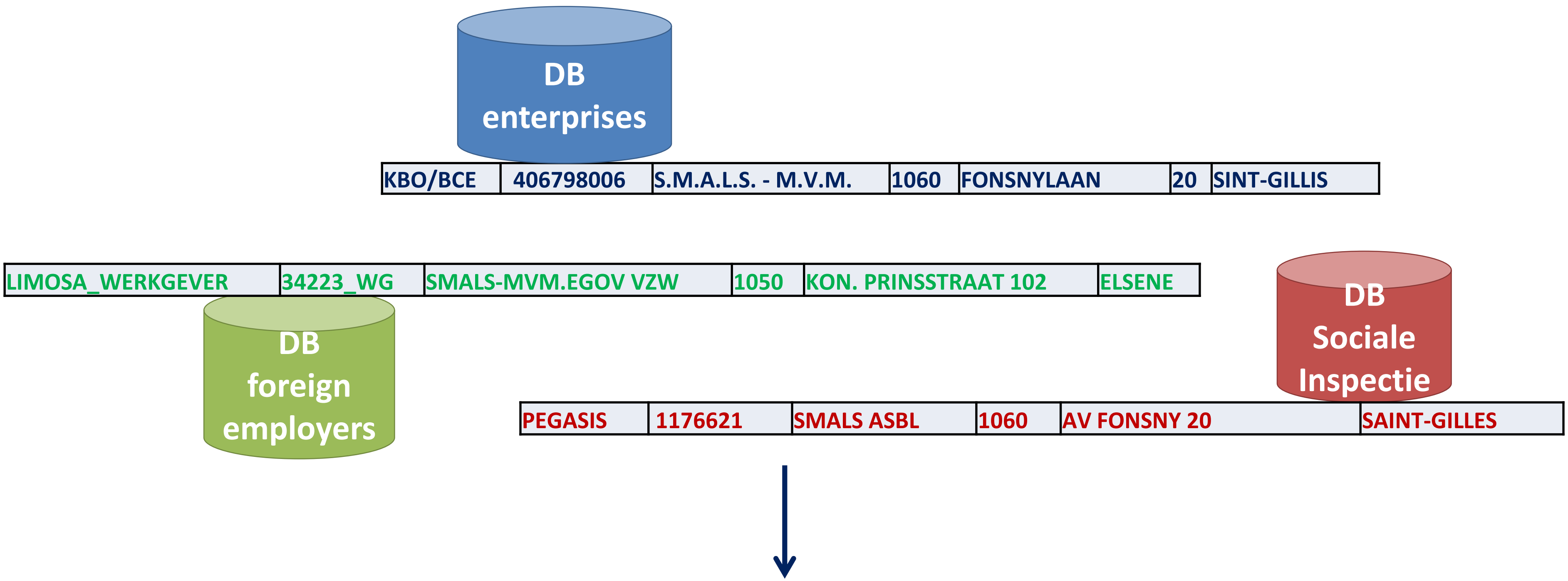
4. Fitness for use: a typical situation

→ **The same entity appears as 100s of different enterprises**



4. Fitness for use: a typical situation

→ The same entity appears as 100s of different enterprises



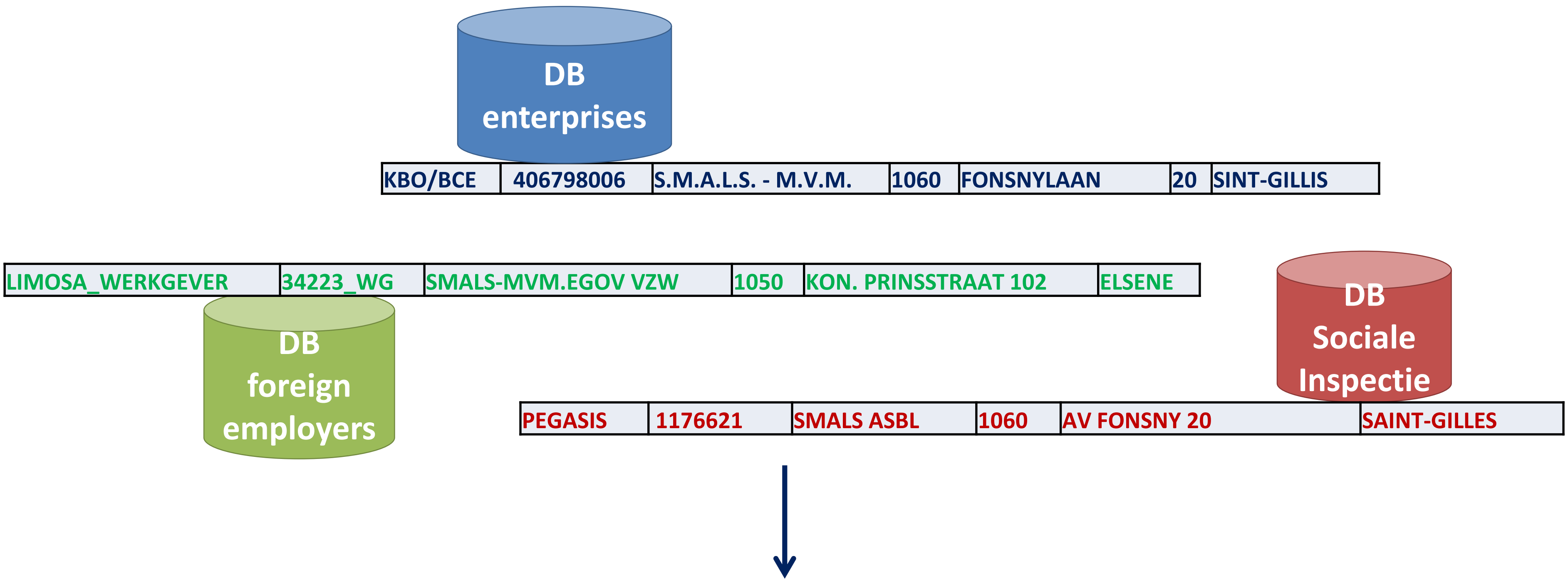
« We need to view data as enterprises, not rows. »

ULB MASTIC en

Smals ICT for society

4. Fitness for use: a typical situation

→ **The same entity appears as 100s of different enterprises**



« We need to view data as enterprises, not rows. »
But how do you do that?...

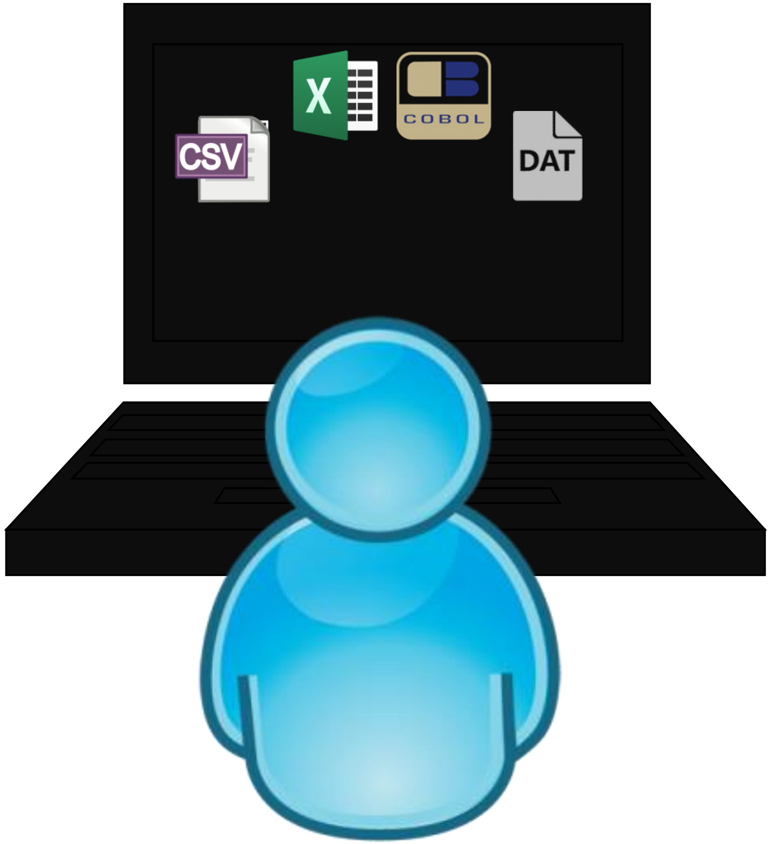
4. Fitness for use

- Frequent use cases
  - Creating a new repertory from external sources
  - Integration of IT systems and DBs
  - Fusions and migrations between administrations
  - Predictive analytics and statistical modeling
  - Etc.

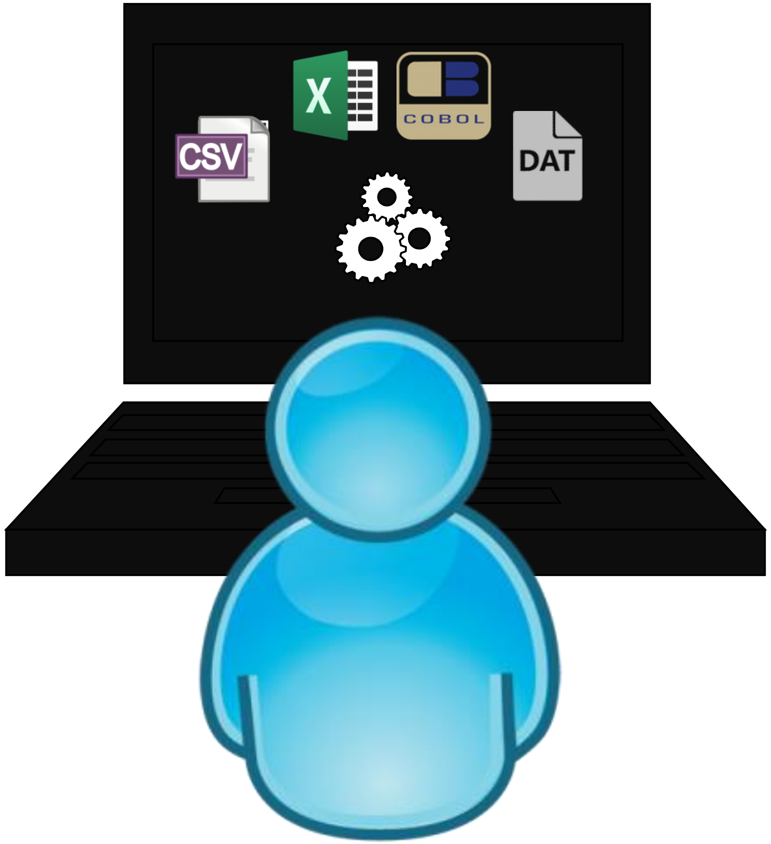
- Important financial impact in Belgium (social security)
  - € 65 billion / year
- …and elsewhere
  - « $3,1 Trillions/year in the US, which is about 20 percent of the Gross Domestic Product. » - Redman T., *Getting in front on data*, Technics Publications, Denville (New Jersey, USA), 2016

Contents

4. How DQ tools work: locally



4. How DQ tools work: locally

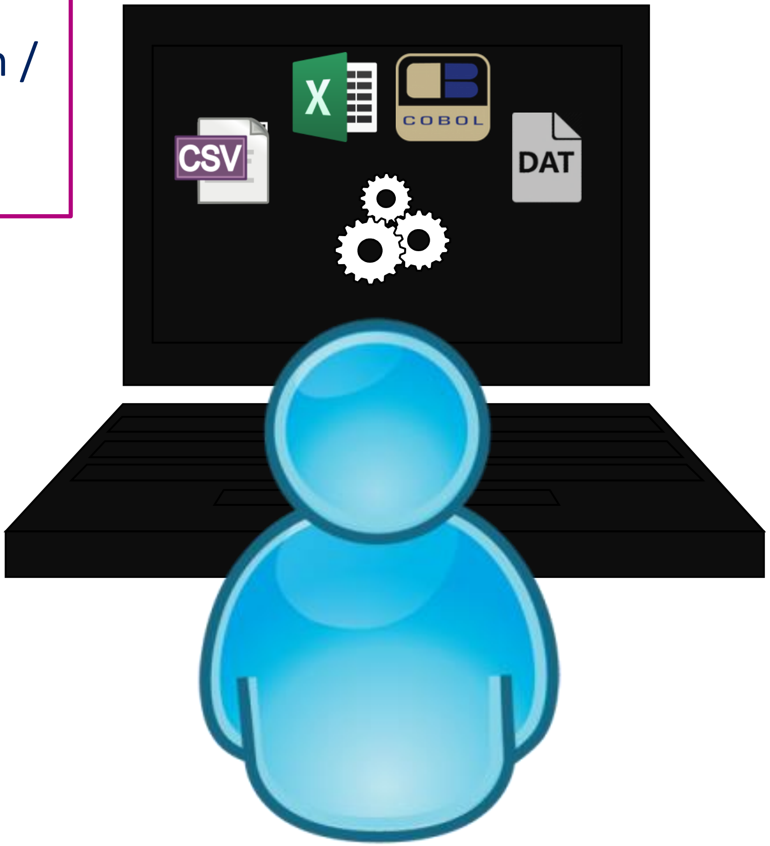


NB : some tools
do offer DB connectivity,
mostly read-only,
in addition to
the common file
import/export
capabilities
(e.g. OpenRefine)

ULB MA en STIC
Smals ICT for society

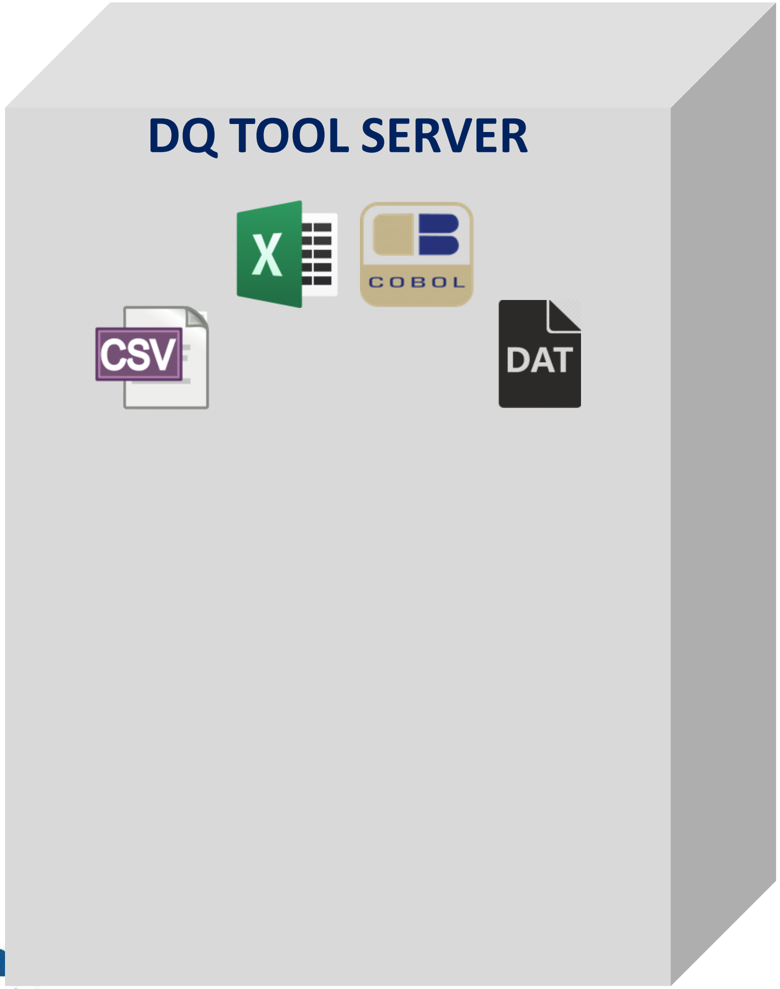
4. How DQ tools work: locally

E.g. : OpenRefine,
Trifacta Wrangler,
Talend Data Preparation / Open Studio,
etc.

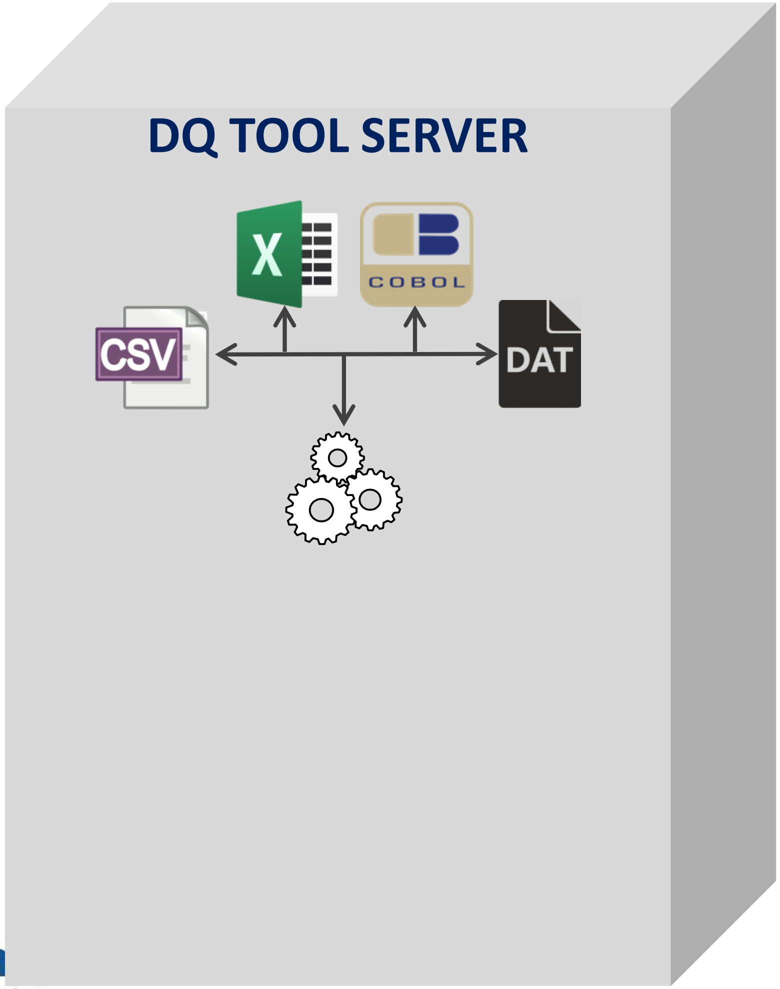


NB : some tools do offer DB connectivity, mostly read-only, in addition to the common file import/export capabilities (e.g. OpenRefine)

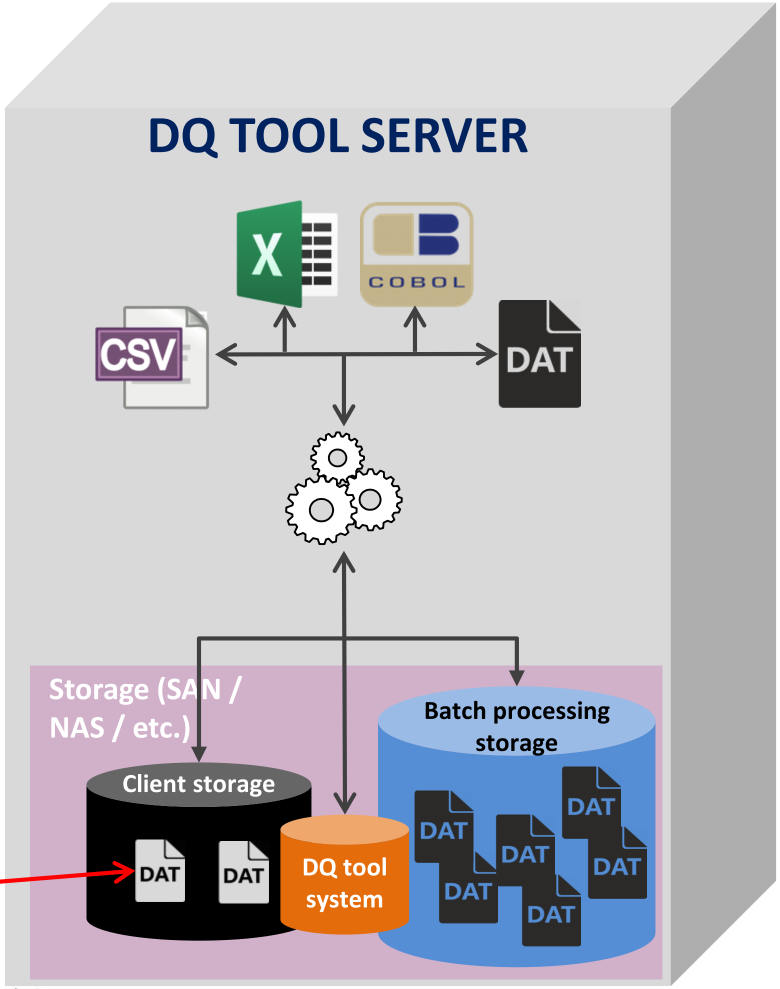
4. How DQ tools work: client-server / multitier



4. How DQ tools work: client-server / multitier



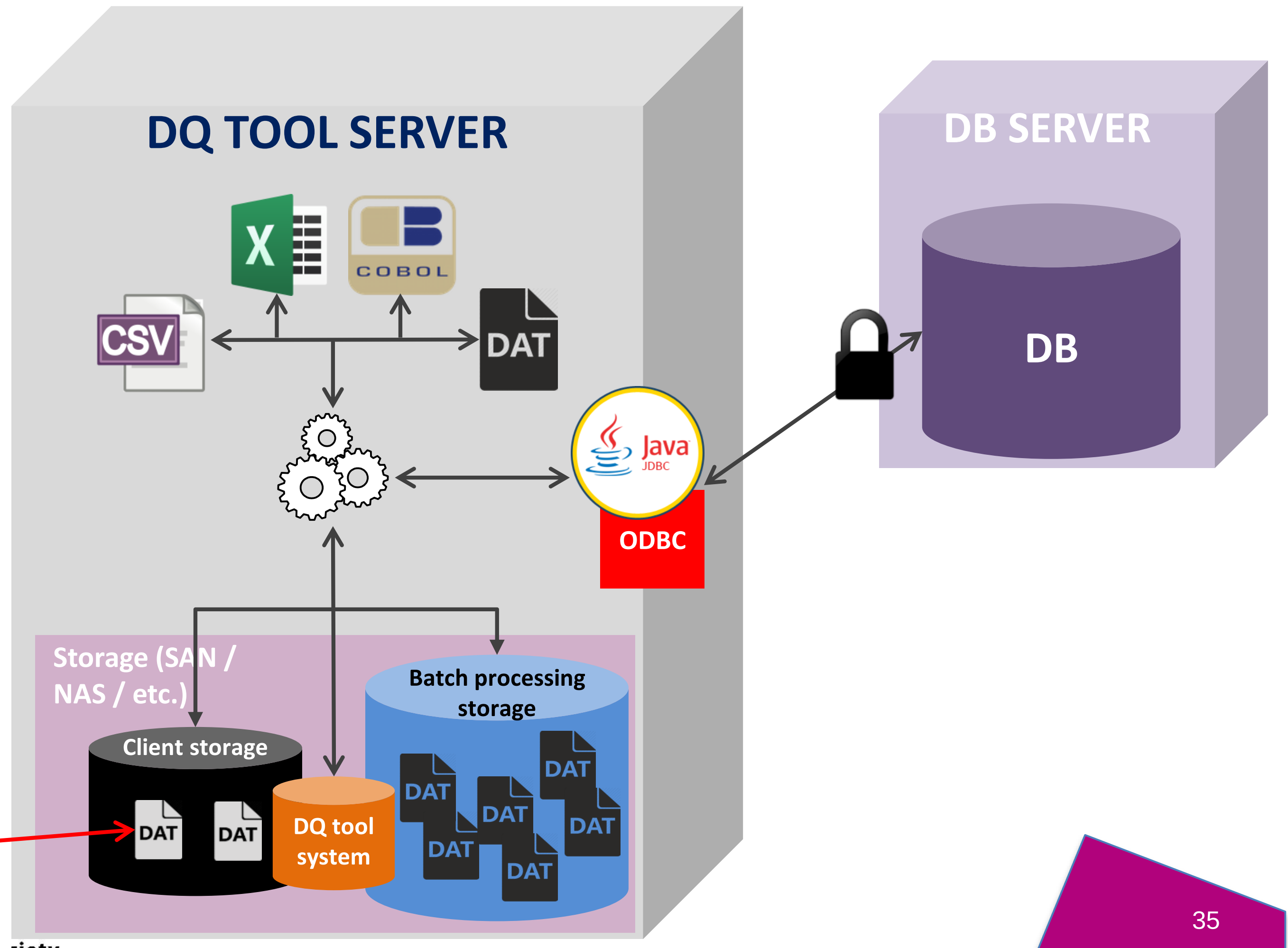
4. How DQ tools work: client-server / multitier



Open or proprietary file format

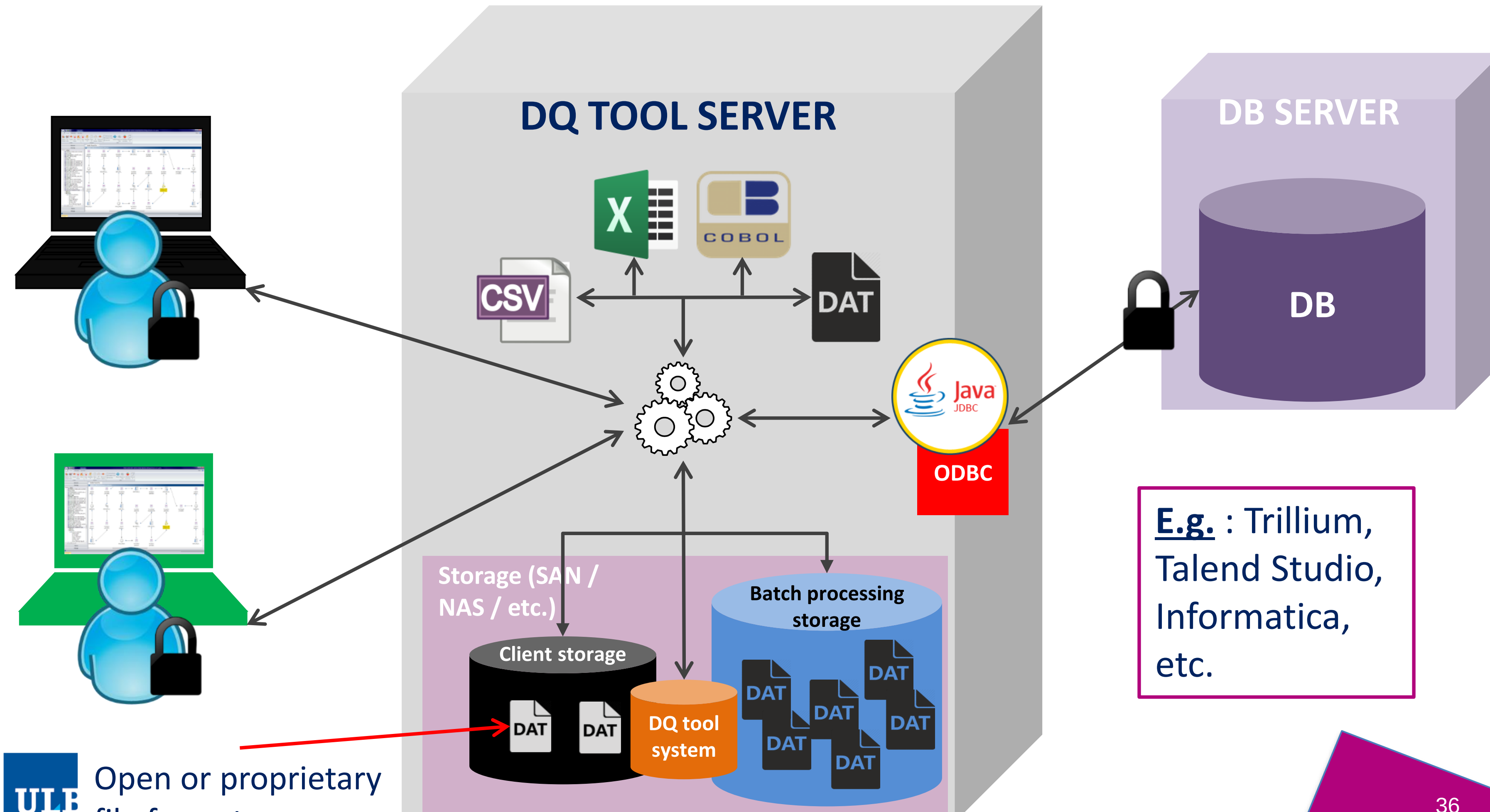
ciety

4. How DQ tools work: client-server / multitier



Open or proprietary file format

4. How DQ tools work: client-server / multitier



E.g. : Trillium, Talend Studio, Informatica, etc.

Contents

Contents

Contents

ULB MASTIC Smals ICT for society

Contents

1. Profiling: Main concepts

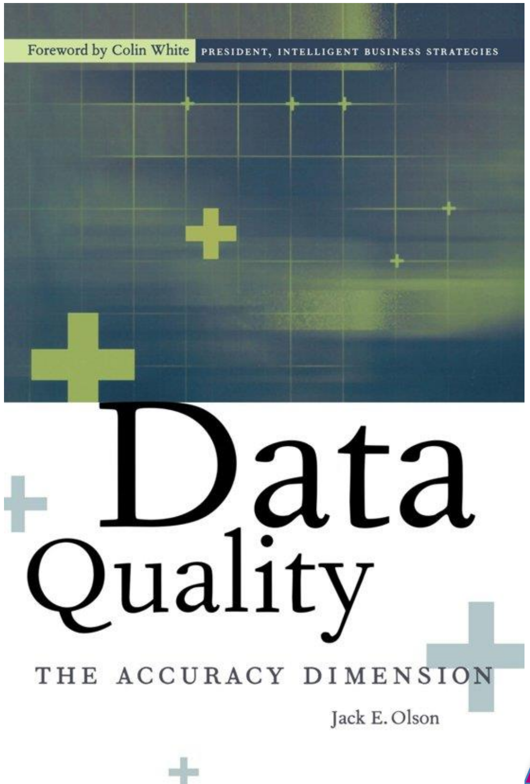
1. What is it?

2. Profiling with a DQ tool

1.1. Main concepts: what is profiling?

« The use of analytical techniques to discover the (…) structure, content and quality of data. »

Olson J., *Data Quality: the Accuracy Dimension*. Elsevier: The Morgan-Kaufmann Series in Database Management, 2002.



1.1. Main concepts: what is profiling?

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

Metadata
(inaccurate, incomplete)

project
Data Profiling

Metadata
(corrected)

Real data
(complete, quality=?)

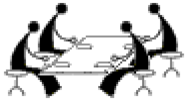
DQTools

Facts about data
(quality = !)

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

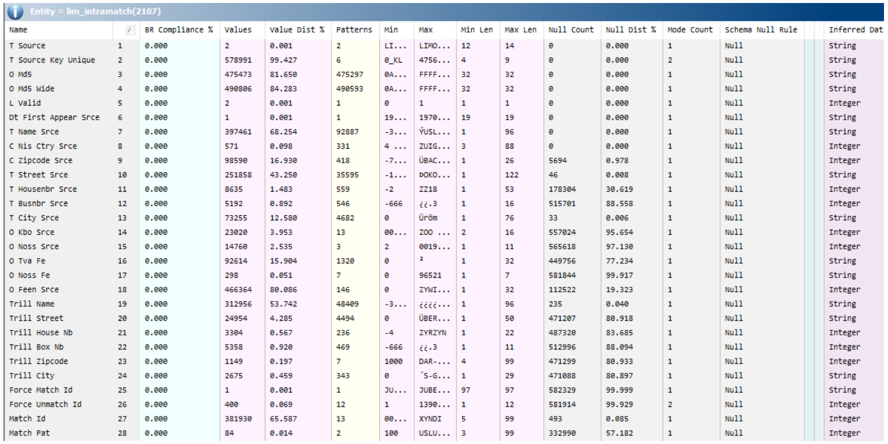
Business people
DQ Analyst team



Source: Y. Bontemps, D. Van Dromme

1.2. Profiling with a Data Quality Tool

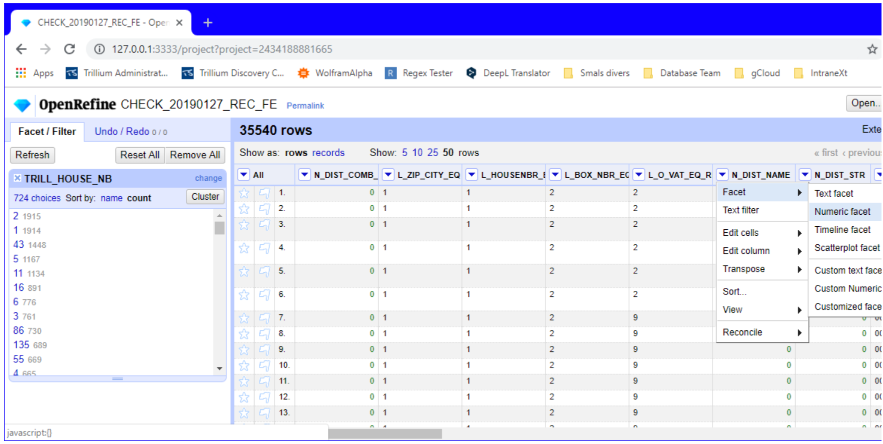
- Automatic analysis upon data loading or « you seek it »



Entity = lim_intramatch(2107)

Name		BR Compliance %	Values	Value Dist %	Patterns	Min	Max	Min Len	Max Len	Null Count	Null Dist %	Mode Count	Schema Null Rule	Inferred Dat
T Source	1	0.000	2	0.001	2	LI...	LIMO...	12	14	0	0.000	1	Null	String
T Source Key Unique	2	0.000	578991	99.427	6	0_KL	4756...	4	9	0	0.000	2	Null	String
O Md5	3	0.000	475473	81.650	475297	0A...	FFFF...	32	32	0	0.000	1	Null	String
O Md5 Wide	4	0.000	490806	84.283	490593	0A...	FFFF...	32	32	0	0.000	1	Null	String
L Valid	5	0.000	2	0.001	1	0	1	1	1	0	0.000	1	Null	Integer
Dt First Appear Srce	6	0.000	1	0.001	1	19...	1970...	19	19	0	0.000	1	Null	String
T Name Srce	7	0.000	397461	68.254	92887	-3...	ÝUSL...	1	96	0	0.000	1	Null	String
C Nis Ctry Srce	8	0.000	571	0.098	331	4 ...	ZUIG...	3	88	0	0.000	1	Null	Integer
C Zipcode Srce	9	0.000	98590	16.930	418	-7...	ÜBAC...	1	26	5694	0.978	1	Null	Integer
T Street Srce	10	0.000	251858	43.250	35595	-1...	ÞOKO...	1	122	46	0.008	1	Null	String
T Housenbr Srce	11	0.000	8635	1.483	559	-2	ZZ18	1	53	178304	30.619	1	Null	Integer
T Busnbr Srce	12	0.000	5192	0.892	546	-666	¿¿.3	1	16	515701	88.558	1	Null	Integer
T City Srce	13	0.000	73255	12.580	4682	0	Üröm	1	76	33	0.006	1	Null	String
O Kbo Srce	14	0.000	23020	3.953	13	00...	ZOO ...	2	16	557024	95.654	1	Null	Integer
O Noss Srce	15	0.000	14760	2.535	3	2	0019...	1	11	565618	97.130	1	Null	Integer
O Tva Fe	16	0.000	92614	15.904	1320	0	²	1	32	449756	77.234	1	Null	String
O Noss Fe	17	0.000	298	0.051	7	0	96521	1	7	581844	99.917	1	Null	String
O Feen Srce	18	0.000	466364	80.086	146	0	ZYWI...	1	32	112522	19.323	1	Null	Integer
Trill Name	19	0.000	312956	53.742	48409	-3...	¿¿¿¿...	1	96	235	0.040	1	Null	String
Trill Street	20	0.000	24954	4.285	4494	0	ÜBER...	1	50	471207	80.918	1	Null	String
Trill House Nb	21	0.000	3304	0.567	236	-4	ZYRZYN	1	22	487320	83.685	1	Null	Integer
Trill Box Nb	22	0.000	5358	0.920	469	-666	¿¿.3	1	11	512996	88.094	1	Null	Integer
Trill Zipcode	23	0.000	1149	0.197	7	1000	DAR-...	4	99	471299	80.933	1	Null	Integer
Trill City	24	0.000	2675	0.459	343	0	´S-G...	1	29	471088	80.897	1	Null	String
Force Match Id	25	0.000	1	0.001	1	JU...	JUBE...	97	97	582329	99.999	1	Null	String
Force Unmatch Id	26	0.000	400	0.069	12	1	1390...	1	12	581914	99.929	2	Null	Integer
Match Id	27	0.000	381930	65.587	13	00...	XYNDI	5	99	493	0.085	1	Null	Integer
Match Pat	28	0.000	84	0.014	2	100	USLU...	3	99	332990	57.182	1	Null	Integer

or



- Concretely, data about your data
  - « **metadata** »
  - Quantitative and qualitative
  - Fundamental and formal
  - != BI ; main focus = DQ

1.2. Profiling with a Data Quality Tool

- At the dataset level
  - **Entity / table level profiling**

1.2. Profiling with a Data Quality Tool

- At the dataset level
  - **Entity / table level profiling**
- Field per field
  - **Attribute / column level profiling**

1.2. Profiling with a Data Quality Tool

- At the dataset level
  - **Entity / table level profiling**
- Field per field
  - **Attribute / column level profiling**
- Relations into the data
  - Primary **Keys** analysis
  - Functional **Dependency analysis**
  - Referential constraints with **Join analysis**

1.2. Profiling with a Data Quality Tool

- At the dataset level
  - **Entity / table level profiling**
- Field per field
  - **Attribute / column level profiling**
- Relations into the data
  - Primary **Keys** analysis
  - Functional **Dependency analysis**
  - Referential constraints with **Join analysis**
- Consistency and business logic
  - **Business rules** analysis

Contents

ULB MAen STIC

Smals ICT for society

2. Drill-down into entities and attributes

- Browsing through data at various levels
  - Zomming on a specific observation : drill-down
  - Going back one layer : drill-up

- Full path : from entity / table metadata, through intermediary measures, to data

2. Drill-down into entities and attributes

1. Entity / table level profiling

2. Attribute / column level profiling

2.1. Entity / table level profiling

- Summary about entity business rules

Business Rules	1	The number of business rules defined for this entity
Enabled Business Rules	1	The number of enabled business rules
Passing Business Rules	0	The number of passing business rules
Failing Business Rules	1	The number of failing business rules

2.1. Entity / table level profiling

- Summary about entity business rules

Business Rules	1	The number of business rules defined for this entity
Enabled Business Rules	1	The number of enabled business rules
Passing Business Rules	0	The number of passing business rules
Failing Business Rules	1	The number of failing business rules

Drill-down

Name	Threshold	Result	Passing...	Created By	Date Created
If country = Belgium then Postc...	100	failed	98.044	ganha	2018/03/01 17:21:51

If country = Belgium
then Postcode ~ NNNN

2.1. Entity / table level profiling

- Summary about entity business rules

Business Rules	1	The number of business rules defined for this entity
Enabled Business Rules	1	The number of enabled business rules
Passing Business Rules	0	The number of passing business rules
Failing Business Rules	1	The number of failing business rules

Drill-down

Name	Threshold	Result	Passing...	Created By	Date Created
If country = Belgium then Postc...	100	failed	98.044	ganha	2018/03/01 17:21:51

Drill-down to failing rows

If country = Belgium then Postcode ~ NNNN

T Source	T Name Srce	C Nis Ctry Srce	T City Srce	C Zipcode Srce
KBO/BCE	PLOVIE,...	150		9999999999
KBO/BCE	DELWICH...	150		9999999999
KBO/BCE	MOVIES	150	LEURHANDEL	****

2.2. Attribute / column level profiling

- Values counts and distinct measures

Attribute = fim_20171029(64).T Name Srce

Metadata	Value	Description
Value Count	582330	The total count of values in the attribute
Value Dist %	68.255	The measure of how distinct the attribute is
Values	397469	The number of distinct values in the attribute

Drill-down, sort on length descending

Value	Frequency	Dist %	Length
1	1	0.001	1
B	1	0.001	1
D	1	0.001	1
J	1	0.001	1
O	1	0.001	1
x	1	0.001	1
-	4	0.001	1
X	8	0.001	1

2.2. Attribute / column level profiling

- Datatype inference

Name	Strings	Strings Dist %	Decimals	Dec Dist %	Integers	Integer Dist %	Inferred Datatype
C Zipcode Srce	72948	43.579	12	0.002	25630	55.441	Integer
O Tva Fe	80253	20.156	9	0.002	12400	2.537	String
T Busnbr Srce	3104	3.037	47	0.017	2043	8.317	Integer
T Housenbr Srce	4154	3.965	9	0.002	4097	65.379	Integer
T Street Srce	251504	99.646	1	0.001	416	0.346	String

2.2. Attribute / column level profiling

- Datatype inference

Name	Strings	Strings Dist %	Decimals	Dec Dist %	Integers	Integer Dist %	Inferred Datatype
C Zipcode Srce	72948	43.579	12	0.002	25630	55.441	Integer
O Tva Fe	80253	20.156	9	0.002	12400	2.537	String
T Busnbr Srce	3104	3.037	47	0.017	2043	8.317	Integer
T Housenbr Srce	4154	3.965	9	0.002	4097	65.379	Integer
T Street Srce	251504	99.646	1	0.001	416	0.346	String

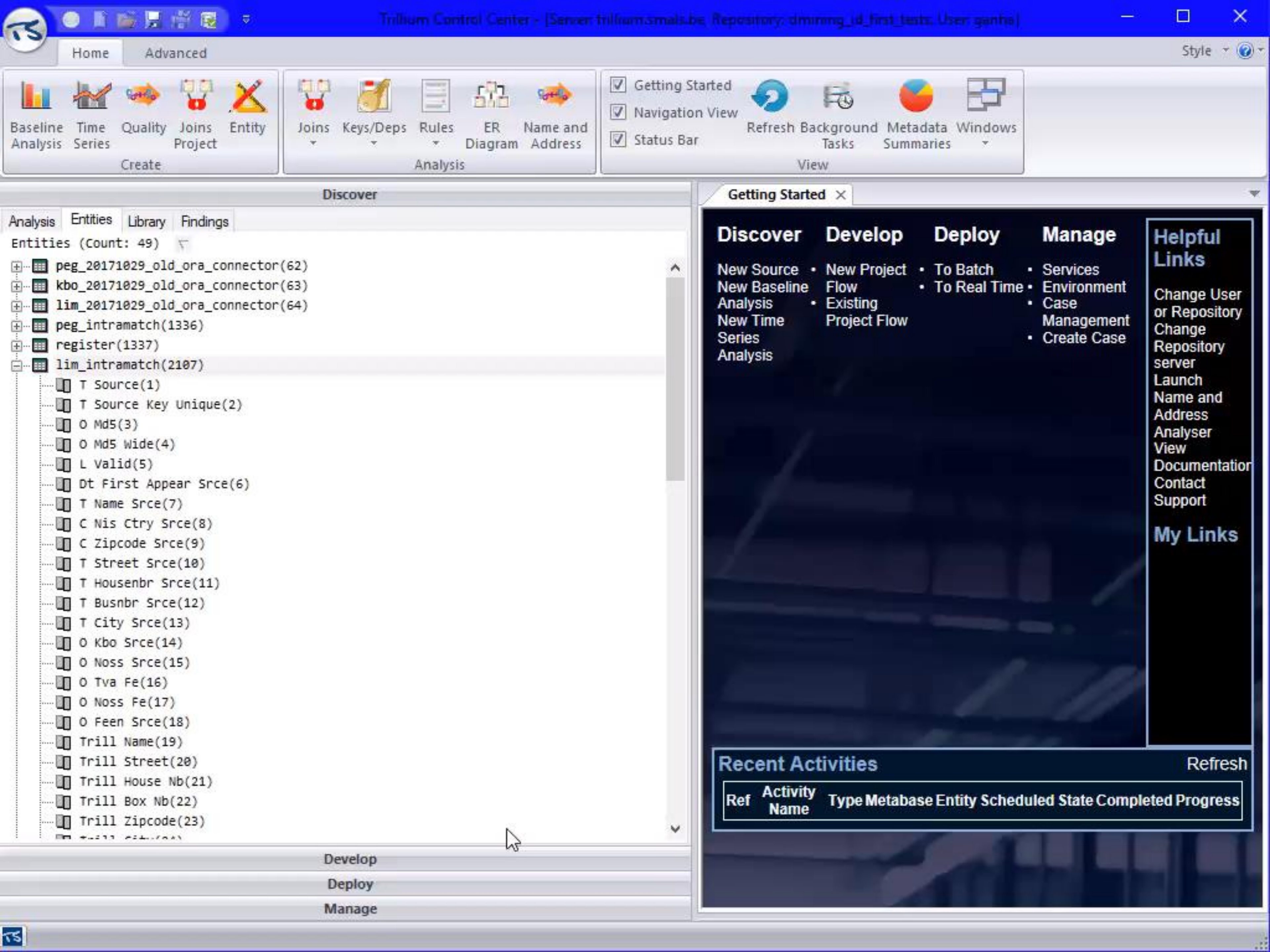
2.2. Attribute / column level profiling

- Datatype inference

Name	Strings	Strings Dist %	Decimals	Dec Dist %	Integers	Integer Dist %	Inferred Datatype
C Zipcode Srce	72948	43.579	12	0.002	25630	55.441	Integer
O Tva Fe	80253	20.156	9	0.002	12400	2.537	String
T Busnbr Srce	3104	3.037	47	0.017	2043	8.317	Integer
T Housenbr Srce	4154	3.965	9	0.002	4097	65.379	Integer
T Street Srce	251504	99.646	1	0.001	416	0.346	String

Drill-down

Value	Frequency	Dist %
Iam not payer	372	0.064
I am not payer VAT	259	0.044
iam not pay vat	45	0.008
I am not responsible to get VAT number.	10	0.002



Trillium Control Center
Home
Advanced
Style
Baseline Analysis
Time Series
Quality
Joins Project
Entity
Create
Joins
Keys/Deps
Rules
ER Diagram
Name and Address
Analysis
Getting Started
Navigation View
Status Bar
Refresh
Background Tasks
Metadata Summaries
Windows
View
Discover
Analysis
Entities
Library
Findings
Entities (Count: 49)
peg_20171029_old_ora_connector(62)
kbo_20171029_old_ora_connector(63)
lim_20171029_old_ora_connector(64)
peg_intramatch(1336)
register(1337)
lim_intramatch(2107)
T Source(1)
T Source Key Unique(2)
O Md5(3)
O Md5 Wide(4)
L Valid(5)
Dt First Appear Srce(6)
T Name Srce(7)
C Nis Ctry Srce(8)
C Zipcode Srce(9)
T Street Srce(10)
T Housenbr Srce(11)
T Busnbr Srce(12)
T City Srce(13)
O Kbo Srce(14)
O Noss Srce(15)
O Tva Fe(16)
O Noss Fe(17)
O Feen Srce(18)
Trill Name(19)
Trill Street(20)
Trill House Nb(21)
Trill Box Nb(22)
Trill Zipcode(23)
Develop
Deploy
Manage
Getting Started
Discover
New Source
New Baseline Analysis
New Time Series Analysis
Develop
New Project Flow
Existing Project Flow
Deploy
To Batch
To Real Time
Manage
Services
Environment
Case Management
Create Case
Helpful Links
Change User or Repository
Change Repository server
Launch Name and Address Analyser View
Documentation
Contact Support
My Links
Recent Activities
Refresh
Ref
Activity Name
Type
Metabase
Entity
Scheduled
State
Completed
Progress

Contents

3. Relations into the data

1. Keys analysis

   Discover primary key candidates or check their validity

2. Dependencies analysis

   Discover or check functional dependencies

3. Joins analysis

   Check referential integrity and foreign keys

3.1. Keys analysis

- Looking for highly unique values

- In specific tools : while loading data, « keys discovery »
  - Analyzing a sample of rows (e.g. 10 000 rows)
  - Looking for atomic or composite (e.g. 2 attributes) keys
  - Keeping candidates that are above a certain uniqueness threshold
    - E.g. >= 98% unique

3.1. Keys analysis

Entity = lim_20171029_old_ora_connector(64)

Metadata	Value	Description
Encoding	AL32UTF8	Encoding used when loading the data
Permanent Keys	0	The number of permanent keys for this entity
Discovered Keys	5	The number of discovered keys for this entity
Permanent Deps	2	The number of permanent dependencies identified for

3.1. Keys analysis

Entity = lim_20171029_old_ora_connector(64)

Metadata	Value	Description
Encoding	AL32UTF8	Encoding used when loading the data
Permanent Keys	0	The number of permanent keys for this entity
Discovered Keys	5	The number of discovered keys for this entity
Permanent Deps	2	The number of permanent dependencies identified for

Drill-down

Entity = lim_20171029_old_ora_connector(64)

Lh Attrs	Status	Verified	Ref	Quality %	Keys	Duplicate Keys	Duplicate Rows
T Source Key Unique	Discovered	No	8	100.000	10000		
O Md5,T City Srce	Discovered	No	8	98.470	9712	135	288
O Md5,T Name Srce	Discovered	No	8	98.120	9648	164	352
O Md5 Wide,T City Srce	Discovered	No	8	98.470	9712	135	288
O Md5 Wide,T Name Srce	Discovered	No	8	98.120	9648	164	352

3.1. Keys analysis

Entity = lim_20171029_old_ora_connector(64)

Metadata	Value	Description
Encoding	AL32UTF8	Encoding used when loading the data
Permanent Keys	0	The number of permanent keys for this entity
Discovered Keys	5	The number of discovered keys for this entity
Permanent Deps	2	The number of permanent dependencies identified for

Entity = lim_20171029_old_ora_connector(64)

Lh Attrs	Status	Verified	Ref	Quality %	Keys	Duplicate Keys	Duplicate Rows
T Source Key Unique	Discovered	No	8	100.000	10000		
O Md5,T City Srce	Discovered	No	8	98.470	9712	135	288
O Md5,T Name Srce	Discovered	No	8	98.120	9648	164	352
O Md5 Wide,T City Srce	Discovered	No	8	98.470	9712	135	288
O Md5 Wide,T Name Srce	Discovered	No	8	98.120	9648	164	352

Is this duplication normal ? Application area specialists need to investigate.

...

Drill-down

Entity lim_20171029_old_ora_connector(64) Key O Md5,T City Srce

O Md5	T City Srce	Dt First...	T ...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	De...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	de...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	de...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	De...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	De...
D5EA082770590CEB7D832DF7CB14B823	Bruxelles	01-JAN-07	DE...

3.2. Functional dependencies analysis

- One or more columns determine the value of others
  - Left1 + Left2 → Right
  - Street + Postcode + City → Country

3.2. Functional dependencies analysis

- One or more columns determine the value of others
  - Left1 + Left2 → Right
  - Street + Postcode + City → Country

- Checking if an expected dependency is met
  - Doubts if unnormalized data model
  - Analytic datasets (denormalized on purpose)
  - Pure data-level issues

- Discover unexpected dependencies
  - Issues in the data model

3.2. Functional dependencies analysis

- One or more columns determine the value of others
  - Left1 + Left2 → Right
  - Street + Postcode + City → Country
- Checking if an expected dependency is met
  - Doubts if unnormalized data model
  - Analytic datasets (denormalized on purpose)
  - Pure data-level issues
- Discover unexpected dependencies
  - Issues in the data model
- Drill down to conflicting values and rows

3.2. Functional dependencies analysis: results of a specific analysis

Entity = kbo_20171029(63)

Lh Attrs	Rh Attr	Quality %	Conflicting LH Values	Conflicting Rows	Verified Date
C Zipcode Srce,T City Srce	C Nis Ctry Srce	99.858	1402	3393	2018/03/19 14:44:07

ULB MASTIC en

Smals ICT for society

3.2. Functional dependencies analysis: results of a specific analysis

Entity = kbo_20171029(63)

Lh Attrs	Rh Attr	Quality %	Conflicting LH Values	Conflicting Rows	Verified Date
C Zipcode Srce,T City Srce	C Nis Ctry Srce	99.858	1402	3393	2018/03/19 14:44:07

Drill-down

Frequency	C Zipcode Srce	T City Srce	C Nis Ctry Srce
1	58636	ISERLOHN	103
1	58636	ISERLOHN	134
7421	3620	LANAKEN	150
7	3620	LANAKEN	999
1	3620	LANAKEN	129
1	VG 1110	TORTOLA	486
1	VG 1110	TORTOLA	112
1	50858	COLOGNE	103
1	50858	COLOGNE	113
1	50858	KOLN	103
1	50858	KOLN	173
1	50858	KOLN	134
29	5001	BELGRADE	150
2	5001	BELGRADE	999

3.2. Functional dependencies analysis: results of a specific analysis

Entity = kbo_20171029(63)

Lh Attrs	Rh Attr	Quality %	Conflicting LH Values	Conflicting Rows	Verified Date
C Zipcode Srce,T City Srce	C Nis Ctry Srce	99.858	1402	3393	2018/03/19 14:44:07

Drill-down

Frequency	C Zipcode Srce	T City Srce	C Nis Ctry Srce
1	58636	ISERLOHN	103
1	58636	ISERLOHN	134
7421	3620	LANAKEN	150
7	3620	LANAKEN	999
1	3620	LANAKEN	129
1	VG 1110	TORTOLA	486
1	VG 1110	TORTOLA	112
1	50858	COLOGNE	103
1	50858	COLOGNE	113
1	50858	KOLN	103
1	50858	KOLN	173
1	50858	KOLN	134
29	5001	BELGRADE	150
2	5001	BELGRADE	999

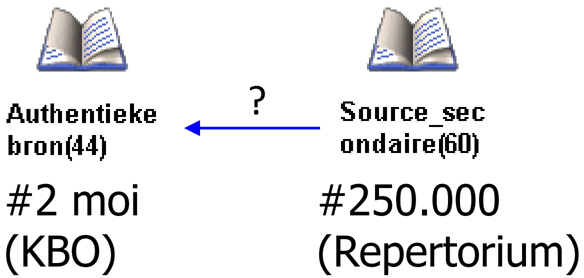
3.3. Joins analysis: referential integrity

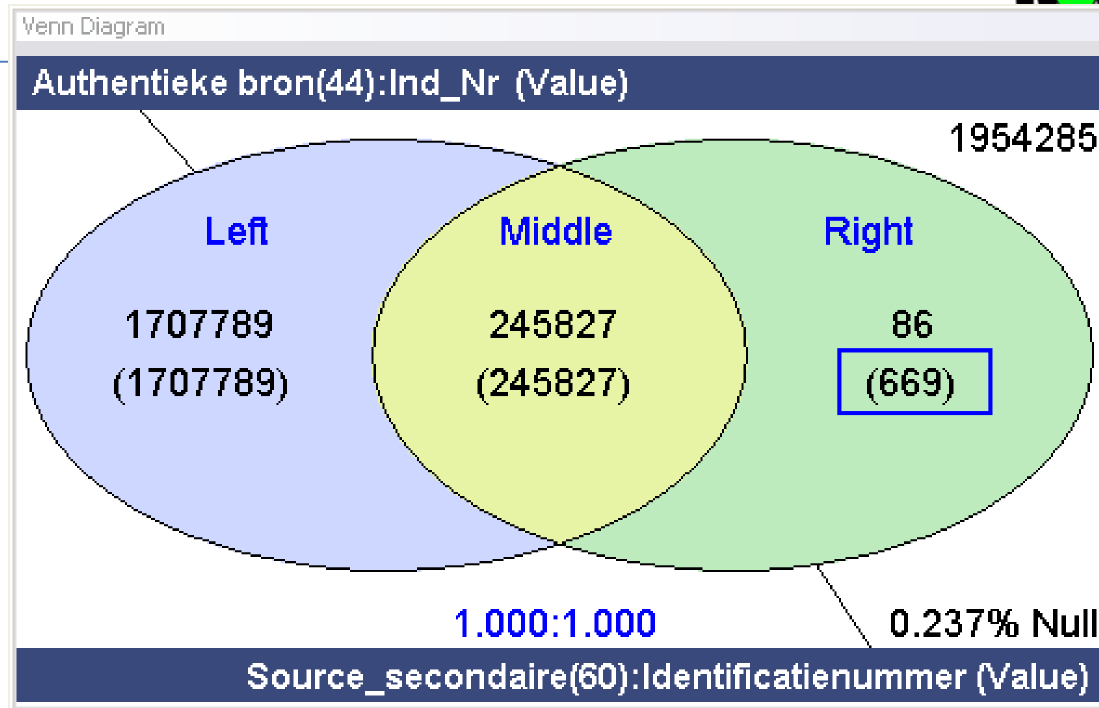
- Join analysis between two or more entities

- Basic principle: Left x Right
  - Metadata on each side
  - Metadata on the intersection
  - Drill-down

- Possible to join on a processed column
  - join(Col)
  - join(ucase(Col))

3.3. Joins analysis: referential integrity





- Source 1: Authentieke bron(44)

  key: « Ind_nr »
- Source 2: Source_secondaire(60)

  fkey: « Identificatienummer »

→ **Join Analysis Source 1 x Source 2**

<u>86 values</u> not found in source 1

(in <u>669 records</u> (so there are doubles))

ULB MASTIC en Smals ICT for society

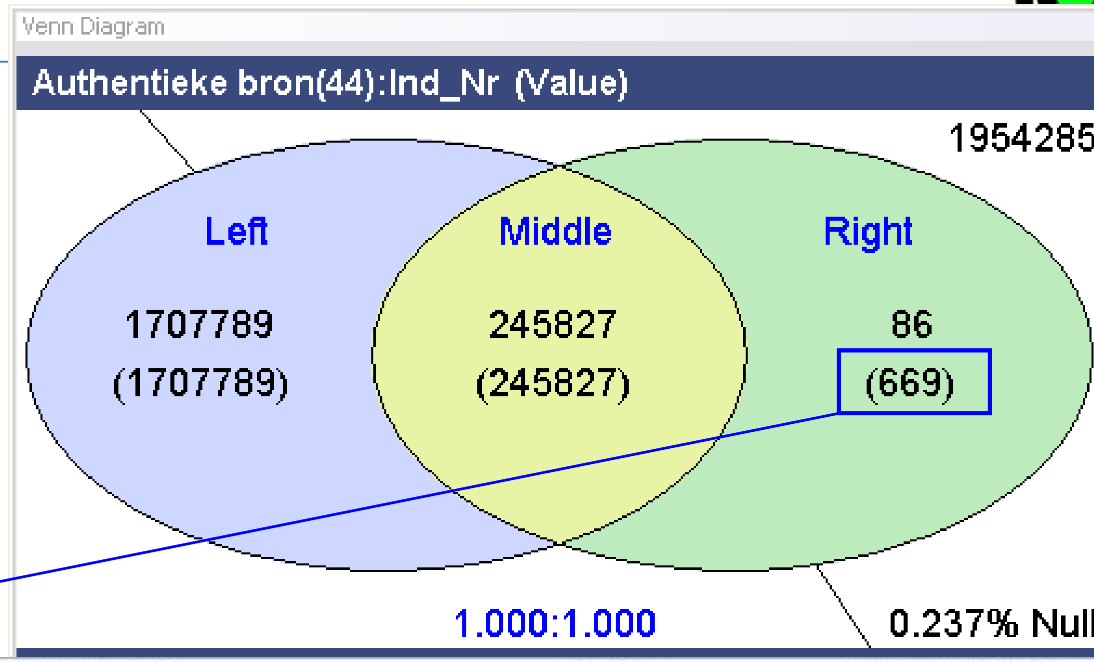
Source: Y. Bontemps, D. Van Dromme

3.3. Joins analysis: referential integrity

Authentieke bron(44) ← ? — **Source_secondaire(60)**

#2 moi (KBO)

#250.000 (Repertorium)



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

Contents

ULB MAenSTIC Smals ICT for society

4. Business rules (BRs)

- 1. Formalizing business constraints
- 2. BR threshold
- 3. Applying BRs
- 4. BRs library

4.1. Formalizing constraints

- Pinpointing business constraints

→ « Postcodes should not contain values other than alphanumeric characters, dashes and spaces. »

4.1. Formalizing constraints

- Pinpointing business constraints

→ « Postcodes should not contain values other than alphanumeric characters, dashes and spaces. »

- Formalizing them:

→ `REGEXP("[^a-zA-Z0-9\- ]",'Postcode') = ""`

4.1. Formalizing constraints

- Pinpointing business constraints

→ « Postcodes should not contain values other than alphanumeric characters, dashes and spaces. »

- Formalizing them:

→ `REGEXP("[^a-zA-Z0-9\- ]",'Postcode') = ""`

- Attribute-specific or inter-attribute

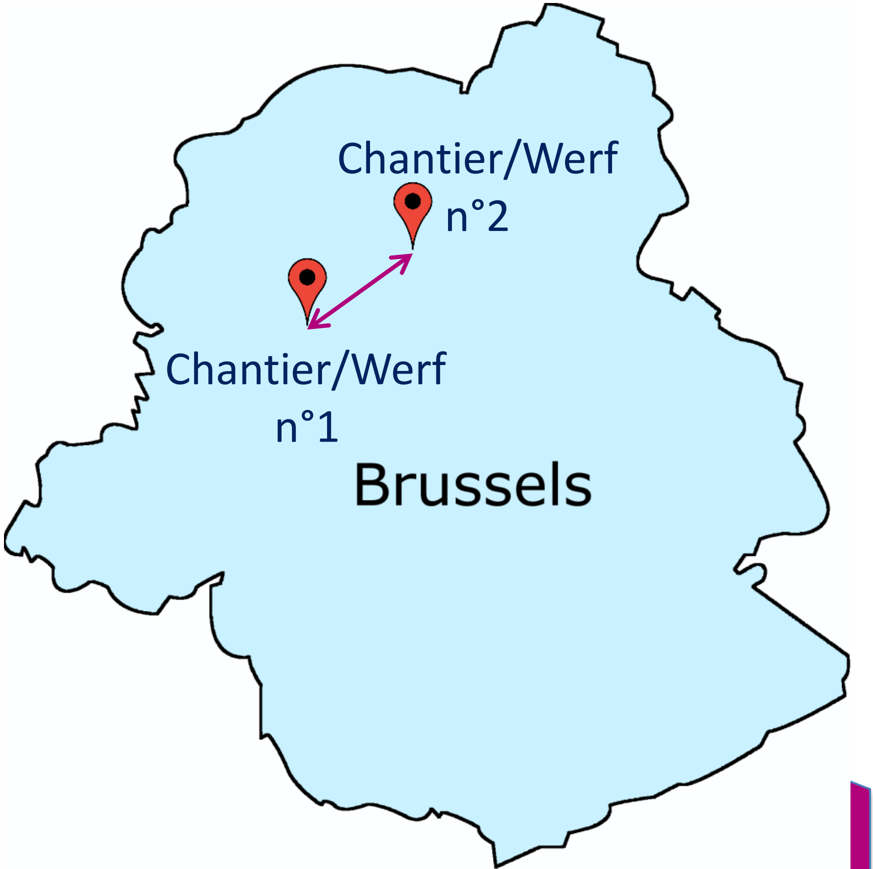
→ `LENGTH(‘Name’) > 3`

→ `LENGTH(‘Name’) > LENGTH(‘Initials’)`

4.1. Formalizing constraints

- **Any string:** names, dates, identifiers, etc.
- E.g.: two work sites declared separately must be more than 100m apart

```
PROXIMITY('lat_1','lat_2',
          'long_1','long_2',
          "KM.000") > "0.100"
```



4.2. Business rules threshold

- Passing threshold T
  - On rows
  - …or on values

4.2. Business rules threshold

- Passing threshold T
  - On rows
  - …or on values
- Example:
  - Rule = `'Name' NOT LIKE "*&*"`
  - Threshold = `50%`

ID	Name
1	IBM
2	STANDARD & POOR'S
3	AXA
2	STANDARD & POOR'S
4	MERRILL LYNCH
2	STANDARD & POOR'S
2	STANDARD & POOR'S
2	STANDARD & POOR'S
2	STANDARD & POOR'S

ULB MASTIC en

Smals ICT for society

4.2. Business rules threshold

- Passing threshold T
  - On rows
  - …or on values

- Example:

  - Rule = `'Name' NOT LIKE "*&*"`
  - Threshold = `50%`
    - On rows: 33% passing < T → **fail**
    - On values: 75% passing >= T → **pass**

ID	Name
1	IBM
2	STANDARD & POOR'S
3	AXA
2	STANDARD & POOR'S
4	MERRILL LYNCH
2	STANDARD & POOR'S
2	STANDARD & POOR'S
2	STANDARD & POOR'S
2	STANDARD & POOR'S

4.3. Applying BRs

Attribute	Name	Threshold	Result	Passing Fraction	Created By	Date Created
T Name Srce	Length more than 3	100	failed	99.192	ganha	2017/11/21 11:0..
T Name Srce	No HTML patterns	100	failed	99.999	ganha	2017/11/21 13:4..
T Name Srce	No special character	100	failed	90.227	ganha	2017/11/21 11:0..
T Name Srce	Not only num	100	failed	99.743	ganha	2017/11/21 17:0..

BRs can **run individually or as sets**

4.3. Applying BRs

Attribute	Name	Threshold	Result	Passing Fraction	Created By	Date Created
T Name Srce	Length more than 3	100	failed	99.192	ganha	2017/11/21 11:0..
T Name Srce	No HTML patterns	100	failed	99.999	ganha	2017/11/21 13:4..
T Name Srce	No special character	100	failed	90.227	ganha	2017/11/21 11:0..
T Name Srce	Not only num	100	failed	99.743	ganha	2017/11/21 17:0..

BRs can **run individually or as sets**

A threshold can be used to allow a **margin of tolerance**, which is 0% here

4.3. Applying BRs

Attribute	Name	Threshold	Result	Passing Fraction	Created By	Date Created
T Name Srce	Length more than 3	100	failed	99.192	ganha	2017/11/21 11:0..
T Name Srce	No HTML patterns	100	failed	99.999	ganha	2017/11/21 13:4..
T Name Srce	No special character	100	failed	90.227	ganha	2017/11/21 11:0..
T Name Srce	Not only num	100	failed	99.743	ganha	2017/11/21 17:0..

BRs can **run individually or as sets**

A threshold can be used to allow a **margin of tolerance**, which is 0% here

Measured results

4.3. Applying BRs

Attribute	Name	Threshold	Result	Passing Fraction	Created By	Date Created
T Name Srce	Length more than 3	100	failed	99.192	ganha	2017/11/21 11:0..
T Name Srce	No HTML patterns	100	failed	99.999	ganha	2017/11/21 13:4..
T Name Srce	No special character	100	failed	90.227	ganha	2017/11/21 11:0..
T Name Srce	Not only num	100	failed	99.743	ganha	2017/11/21 17:0..

BRs can **run individually or as sets**

A threshold can be used to allow a **margin of tolerance**, which is 0% here

Measured results

Audit info

4.3. Applying BRs

Attribute	Name	Threshold	Result	Passing Fraction	Created By	Date Created
T Name Srce	Length more than 3	100	failed	99.192	ganha	2017/11/21 11:0..
T Name Srce	No HTML patterns	100	failed	99.999	ganha	2017/11/21 13:4..
T Name Srce	No special character	100	failed	90.227	ganha	2017/11/21 11:0..
T Name Srce	Not only num	100	failed	99.743	ganha	2017/11/21 17:0..

Drill-down

Attribute = lim_20171029(64).T Name Srce

Value	Frequency	Dist %	Length
ПК "&#1...	1	0.001	96
Amoi Electr...	1	0.001	32
Amoi Electr...	1	0.001	37
paweł	1	0.001	10
Rogoziński, Kr...	1	0.001	26

Contents

5. Profiling report and iterating with business

- **Iterating with data users and application area specialists is key**

- Interpreting profiling results
  - What is not an issue
  - What is an issue
  - Setting priorities
  - Comparing sources

- Follow-up and monitoring can be supported with a profiling report

ULB MA en STIC Smals ICT for society

Contents

ULB MASTIC Smals ICT for society

Contents

Contents

ULB MA en STIC

Smals ICT for society

1. Standardization: Main concepts

1. What is it?

2. How DQ tools process data

1.1. Main concepts: what is data standardization?

- Building standards: unambiguous conventions for a correct formal representation of data based on simple business rules

1.1. Main concepts: what is data standardization?

- Building standards: unambiguous conventions for a correct formal representation of data based on simple business rules
  - Eg: « All mobile numbers should be represented as:

    +NN NNN NNN NNN

    without /, -, (), and with spaces each 3 chars ».

- Conforming the representation of data to the agreed standard
  - Profiling can help discover standardization issues

1.1. Main concepts: what is data standardization?

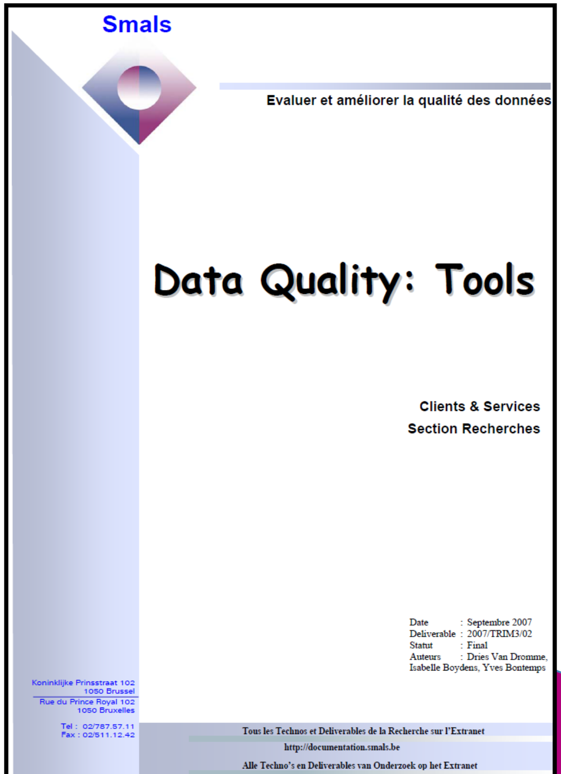
- Building standards: unambiguous conventions for a correct formal representation of data based on simple business rules
  - Eg: « All mobile numbers should be represented as:

    +NN NNN NNN NNN

    without /, -, (), and with spaces each 3 chars ».

- Conforming the representation of data to the agreed standard
  - Profiling can help discover standardization issues

Source: Bontemps Y., Boydens I., Van Dromme D., *Data Quality: tools*, Smals Research, 2007 (http://www.smalsresearch.be/?wpfb_dl=85).



Smals

Evaluer et améliorer la qualité des données

Data Quality: Tools

Clients & Services
Section Recherches

Date : Septembre 2007
Deliverable : 2007/TRIM3/02
Statut : Final
Auteurs : Dries Van Dromme, Isabelle Boydens, Yves Bontemps

Koninklijke Prinsstraat 102 1050 Brussel
Rue du Prince Royal 102 1050 Bruxelles
Tel : 02/787.57.11
Fax : 02/511.12.42

Tous les Technos et Deliverables de la Recherche sur l'Extranet
http://documentation.smals.be
Alle Techno's en Deliverables van Onderzoek op het Extranet

1.1. Main concepts: what is data standardization?

- Solving the lack of standardization *per se*:

1.1. Main concepts: what is data standardization?

- Solving the lack of standardization *per se*:
  - in one data source
  - across databases
    - solving inconsistencies in the (re-)use of data concepts
    - transversal data management, Master Data Management
    - requires breaking down siloes, and governance
  - across institutions
    - Inter-institutional Master Data Management, even more governance

Source: Y. Bontemps, D. Van Dromme

1.1. Main concepts: what is data standardization?

- Solving the lack of standardization *per se*:
  - in one data source
  - across databases
    - solving inconsistencies in the (re-)use of data concepts
    - transversal data management, Master Data Management
    - requires breaking down siloes, and governance
  - across institutions
    - Inter-institutional Master Data Management, even more governance

- Or as an intermediary step in fuzzy matching
  - standardization = best practice
  - greatly improving reliability of matching results

1.2. Main concepts: How DQ tools process data

- Unlike profiling, with standardization (and then matching) :
  - We modify entity / table schemas
    - Create, delete, merge, rename columns
    - Join or split tables
  - We transform the data itself
    - Cleansing, concatenating, splitting…
    - Validating, enriching
    - Merging rows
    - Etc.
- Thus, we need to understand how data gets processed in a DQ tool

1.2. Main concepts: How DQ tools process data – Spreadsheet-like interfaces

- Data is almost permanently shown on-screen
  - Possibly with some statistics
  - Most often, only a sample for performance reasons

- Transformations are done “in-place”
  - Can be recorded as a script for later re-use
  - One final file as a result

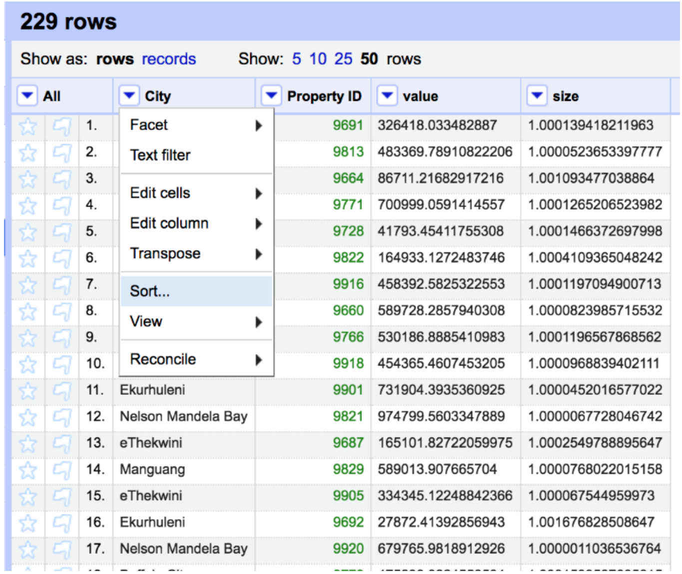
- Lightweight, great for :
  - Quick fixes
  - Reasonable datasets
  - Modest budgets

1.2. Main concepts: How DQ tools process data – Spreadsheet-like interfaces

- Data is almost permanently shown on-screen
  - Possibly with some statistics
  - Most often, only a sample for performance reasons

- Transformations are done "in-place"
  - Can be recorded as a script for later re-use
  - One final file as a result

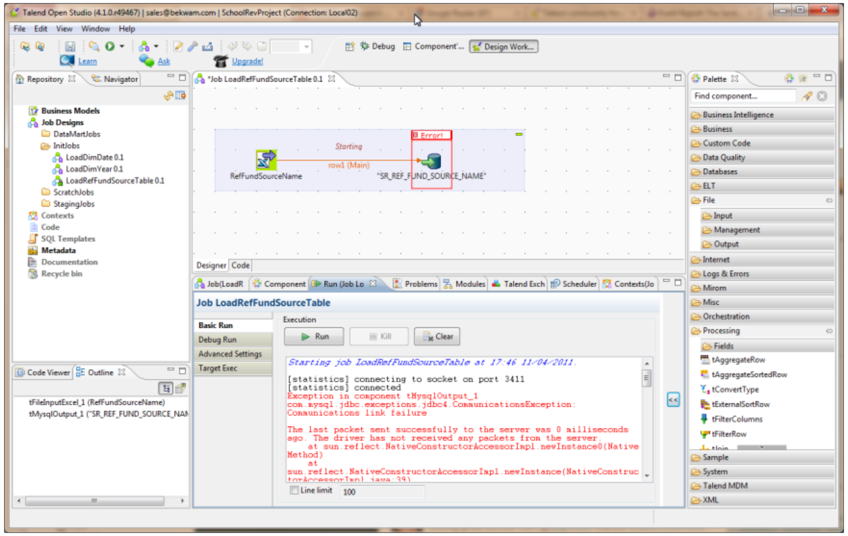
- Lightweight, great for :
  - Quick fixes
  - Reasonable datasets
  - Modest budgets



E.g. : OpenRefine, Trifacta Data Wrangler

1.2. Main concepts: How DQ tools process data – Data Flows / Jobs

- The interface is usually "IDE"-like
  - Central panel : development / data area
  - Side panel(s) : processes palette, entities / tables, projects
  - Bottom panel : console / logs



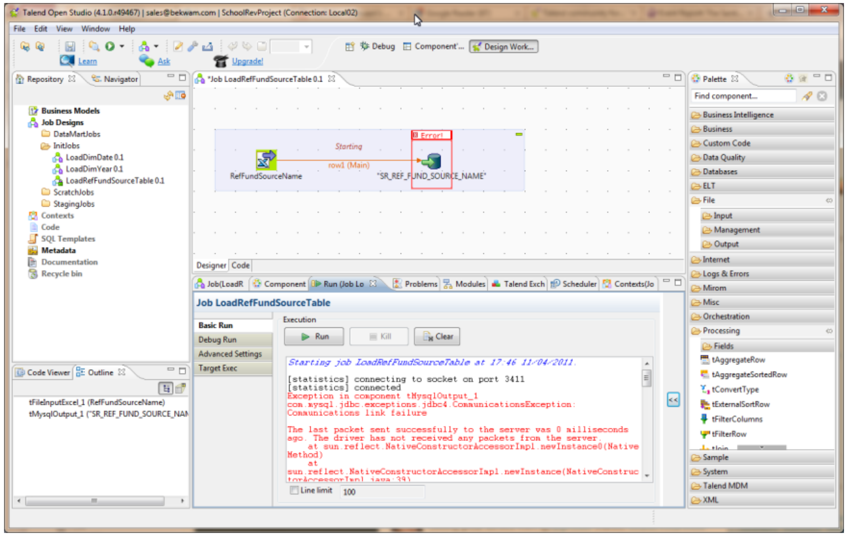
- Data from process(es) to process(es)
  - Input(s) → Process(es) → Output(s)
  - Intermediary files are available
  - Designed to be

- Higher flexibility - Higher complexity
  - Processes are dedicated to specific tasks
  - Each process is a tool by itself
  - Data routing freedom

  → Steeper learning curve

1.2. Main concepts: How DQ tools process data – Data Flows / Jobs

- The interface is usually “IDE”-like
  - Central panel : development / data area
  - Side panel(s) : processes palette, entities / tables, projects
  - Bottom panel : console / logs



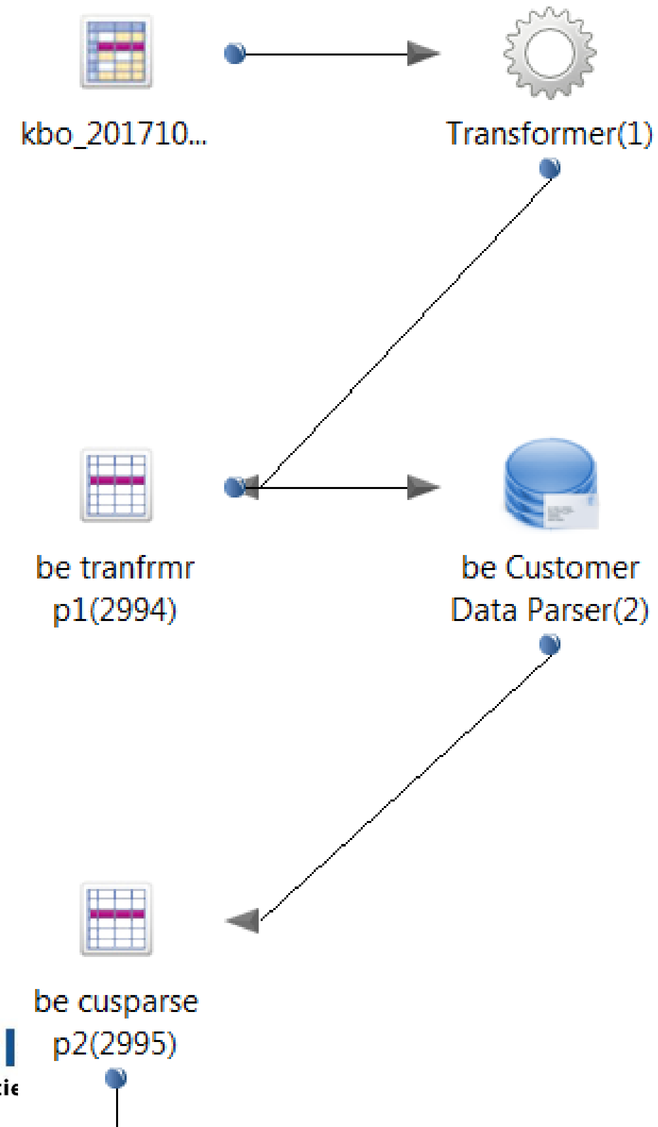
- Data from process(es) to process(es)
  - Input(s) → Process(es) → Output(s)
  - Intermediary files are available
  - Designed to be

- Higher flexibility - Higher complexity
  - Processes are dedicated to specific tasks
  - Each process is a tool by itself
  - Data routing freedom

  → Steeper learning curve

E.g. : Trillium, Talend Studio, Informatica, etc.

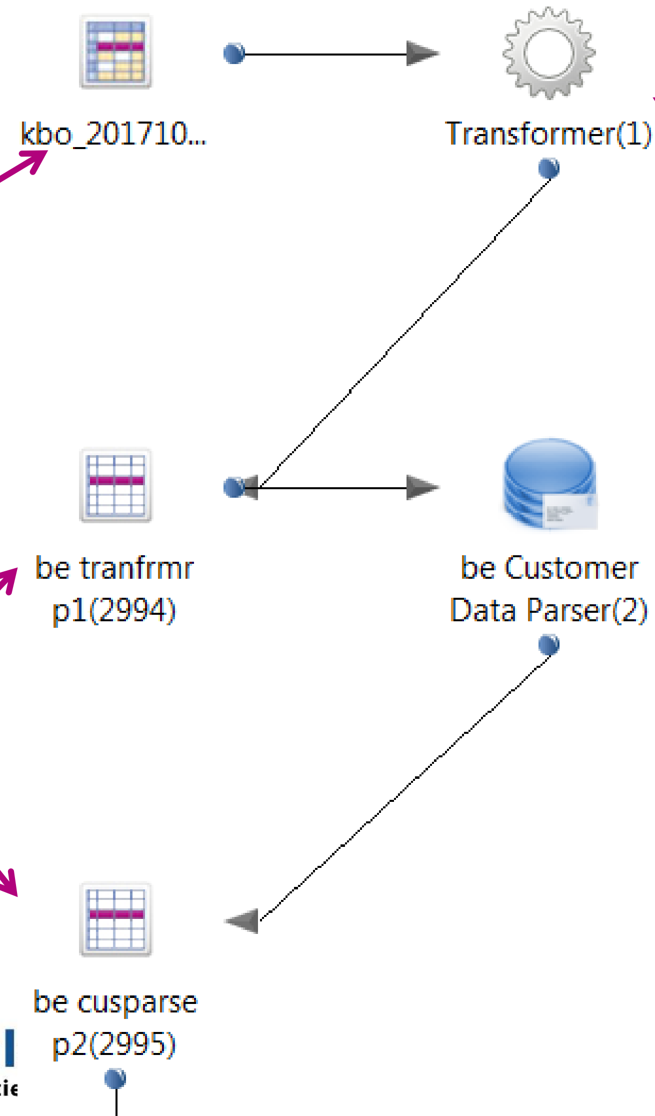
1.2. Main concepts: How DQ tools process data – Data Flows / Jobs



E.g. : Trillium, Talend Studio, Informatica, etc.

1.2. Main concepts: How DQ tools process data – Data Flows / Jobs

Data table (with profiling metadata)



Processes

Data table (no profiling metadata yet)

E.g. : Trillium, Talend Studio, Informatica, etc.

1.2. Main concepts: How DQ tools process data

- As a rule of thumb: Every change in extra attributes
  - Original data never overwritten
  - Comparable and reversible changes

Contents

Introduction: DQ fundamentals

Part 1: Data Profiling

Part 2: Parsing, Standardization & Address enrichment

- **1. Main concepts**
- **2. Conditional operations**
- 3. Parsing-enabled standardization
- 4. Validating and enriching addresses

Part 3: Data matching and Window keys (performance)

Conclusion & questions

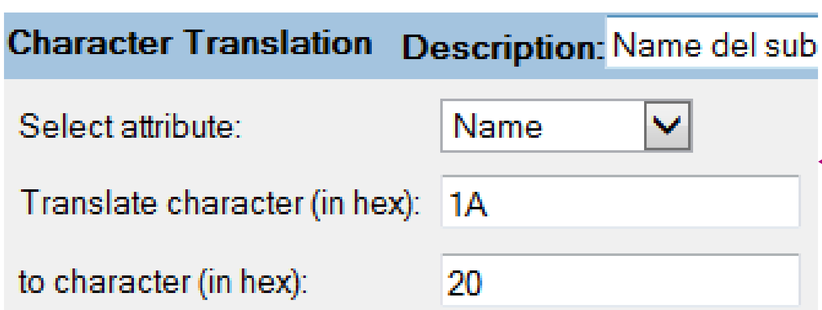
ULB MA en STIC Smals ICT for society

2. Conditional operations

- Built manually
  - Various languages/scripts : Java, GREL, etc.
  - Executed if a condition is met

- **Character-level recodings**
  - Managing invisible characters (e.g. control chars)
  - Untypable, targeted with their hex value
  - E.g. CR LF: 0x0D 0x0A (Windows), Sub: 0x1A (certain OSes)

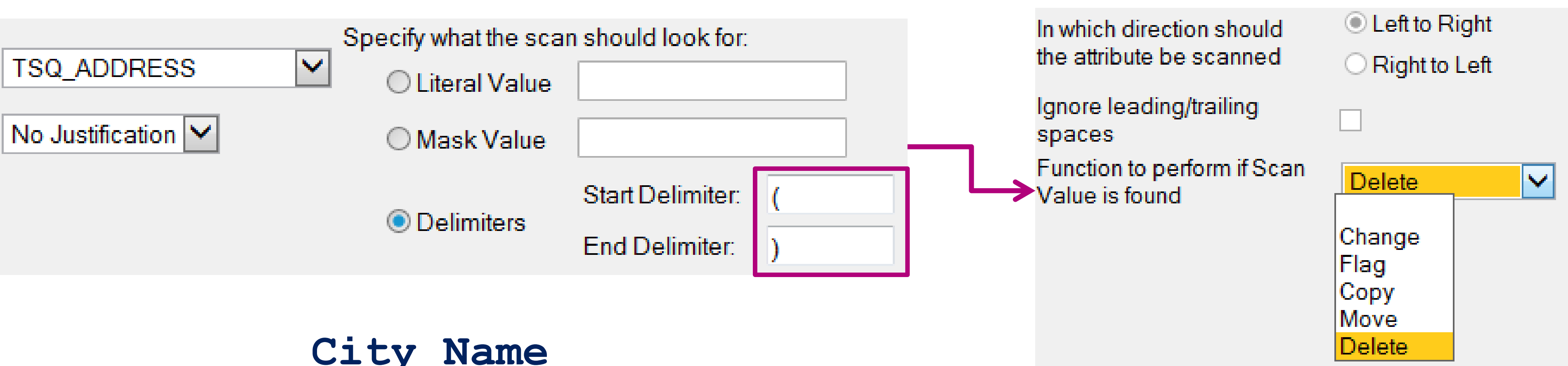
```
"name","street","postal_code","city"
"Smals","Avenue
Fonsny","1060","Bruxelles"
"SmalsSUBSUBSUB","Avenue Fonsny","1060","Bruxelles"
```



2. Conditional operations

- Built manually
  - Various languages/scripts : Java, GREL, etc.
  - Executed if a condition is met

- **In-attribute changes**
  - Moves, substitutions or deletions inside an attribute



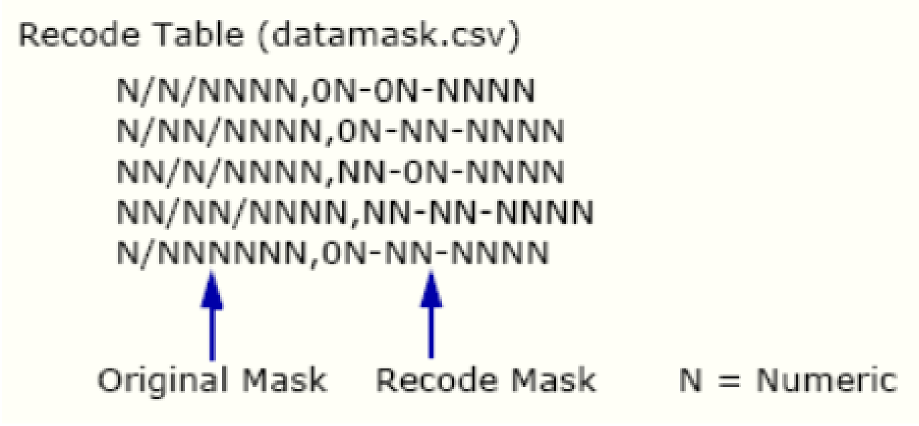
City Name
ANTWERPEN(Mol)

City Name
ANTWERPEN

2. Conditional operations

- Built manually
  - Various languages/scripts : Java, GREL, etc.
  - Executed if a condition is met

- **Join-based recodings**
  - Substitutions, deletions, enrichments, with a « From → To » file



Source : Trillium interactive documentation

2. Conditional operations

- Built manually
  - Various languages/scripts : Java, GREL, etc.
  - Executed if a condition is met

- **Processing / enriching values with webservice calls**
  - E.g. geocoding

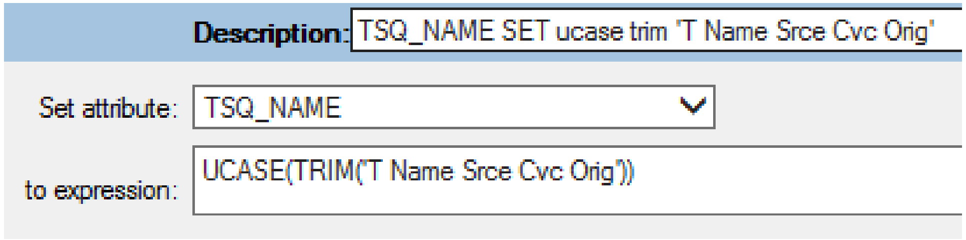
Address	Coordinates
AVENUE FONSNY 20 1060 BRUXELLES	

Address	Coordinates
AVENUE FONSNY 20 1060 BRUXELLES	50.835827;4.3382999

2. Conditional operations

- Built manually
  - Various languages/scripts : Java, GREL, etc.
  - Executed if a condition is met

- **Function-based operations**
  - In- or inter-attribute changes
  - Versatile: Anything that can be the output of a function



Here: filling extra working attribute (TSQ_NAME) with original name value ('T Name Srce') trimmed and fully converted to upper case.

T Name Srce	TSQ_NAME
vzw Smals ASBL	

T Name Srce	TSQ_NAME
vzw Smals ASBL	VZW SMALS ASBL

Contents

ULB MAen STIC Smals ICT for society

3. Parsing-enabled standardization

1. Basic principles

2. Data parsing in a DQ tool

3.1. Parsing-enabled standardization: basic principles

- Processing attributes according to their nature
  - Parsing values into categories
  - Attribute X is a PERSON_NAME
  - Attribute Y is an APPARTMENT_NUMBER

- Knowledge bases
  - Standardization definitions and patterns
  - Specific to each parsing category

- Formal validity of data

3.1. Parsing-enabled standardization: basic principles

FROM	PARSED AS	TO (category-specific rules)
Betty	First name	Elizabeth
Lizzy	First name	Elizabeth
Beth	First name	Elizabeth
asbl	Legal form	ASBL
V. Z.W	Legal form	ASBL
Assoc. Sans But Lucratif	Legal form	ASBL
0485/123 456	Mobile	+32 485 123 456
(0032) 485.123.456	Mobile	+32 485 123 456
+32485123456	Mobile	+32 485 123 456
24/01/17	Date	2017-01-24
Jan. the 24th, 2017	Date	2017-01-24

3.1. Parsing-enabled standardization: basic principles

FROM	PARSED AS	TO (category-specific rules)
Betty Lizzy Beth	First name	Elizabeth
asbl V. Z.W Assoc. Sans But Lucratif	Legal form	ASBL
0485/123 456 (0032) 485.123.456 +32485123456	Mobile	+32 485 123 456
24/01/17 Jan. the 24th, 2017	Date	2017-01-24

IN EXTRA ATTRIBUTES

NOT OVERRIDING ORIGINAL DATA !

3.2. Parsing-enabled standardization: Data parsing in a DQ tool

- Various manual approaches:
  - Conditional operations
  - Regexes

- Some tools go further, providing pre-built, language-specific:
  - Context-free grammar
  - Context-sensitive grammar

  …and the ability to edit / expand them

ULB MA en STIC

Smals ICT for society

3.2. Parsing-enabled standardization: Data parsing in a DQ tool – Grammar-based approach

- Standardize according a set (50 000+) of
  - Rules:

```
'GASTRONOMIE'     NAME    DEF ATT=BUS
```

= If I see « Gastronomie » in a name field, I'll consider it as part of a Business name.

```
'GEERT'           NAME    DEF ATT=GVN-NM1 GEN=M
```

= « Geert » is a first name for a male individual.

```
'POB'             STREET DEF ATT=POBOX REC='POSTBUS'
```

= If I see « POB » in a street attribute, I'll consider that address as a postbox Indicator and I'll replace it with « POSTBUS ».

  - And Patterns:

```
'ALPHA STR-NM HSNO 1ALPHA'
    PATTERN STREET DEF
    REC='STR-NM STR-NM HSNO APT'
```

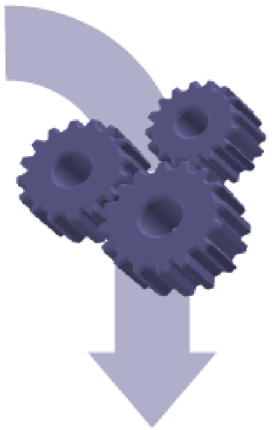
= If I see a pattern made of an alphabetic string, a street name string, a house number and 1 alphabetic character, it's a street and I'll recode it to a street name, house number and appartment.

E.g.: « Fonsny Street 20 B »

ICT for society

3.2. Parsing-enabled standardization: Data parsing in a DQ tool – Product data

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	

Source: Y. Bontemps, D. Van Dromme

Contents

ULB MA en STIC

Smals ICT for society

4. Validating and enriching addresses

1. Basic principles
2. Address validation in a DQ tool
3. Parsing and postal validation hand in hand

3.1. Validating and enriching addresses: basic principles

- Fundamental validity of data
  - != « this looks like a correct address » (parsing)
  - = « this is a correct address »

- Currently, knowledge-based approaches
  - Addresses are very volatile concepts
  - Few standards exist (EU : Inspire ; BE : Best Address)
  - Ubiquitous and strategic problem (clients DBs, public administrations, B2B…)

3.2. Validating and enriching addresses: address validation in a DQ tool

- The local database approach
  - DB stored in the DQ server
    - Provided by the DQ tool editor
    - Theoretically, could be self-provided by the user
  - The server admin will need to push updates to the DB
  - The tool will match input data in batch against this DB

- The webservice approach
  - No local access to the DB itself
  - The service provider pushes updates himself
  - The tool will call an API for each address/in small batches

- Typical results
  - Validating or correcting addresses (or error code if not possible)
  - Statistics about address issues

3.3. Parsing and postal validation hand in hand
Let’s try it !

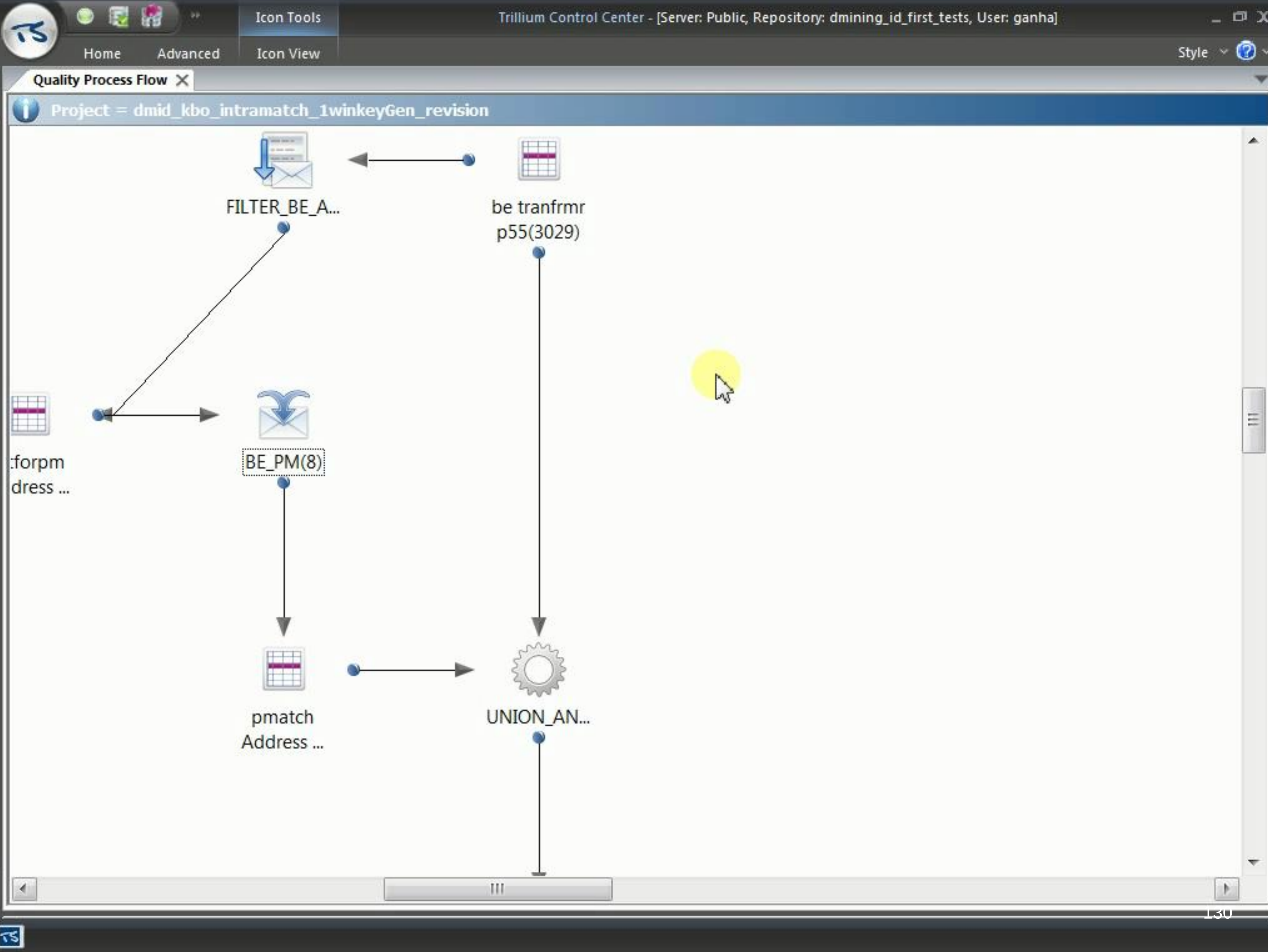
3.3. Parsing and postal validation hand in hand
Let’s try it !

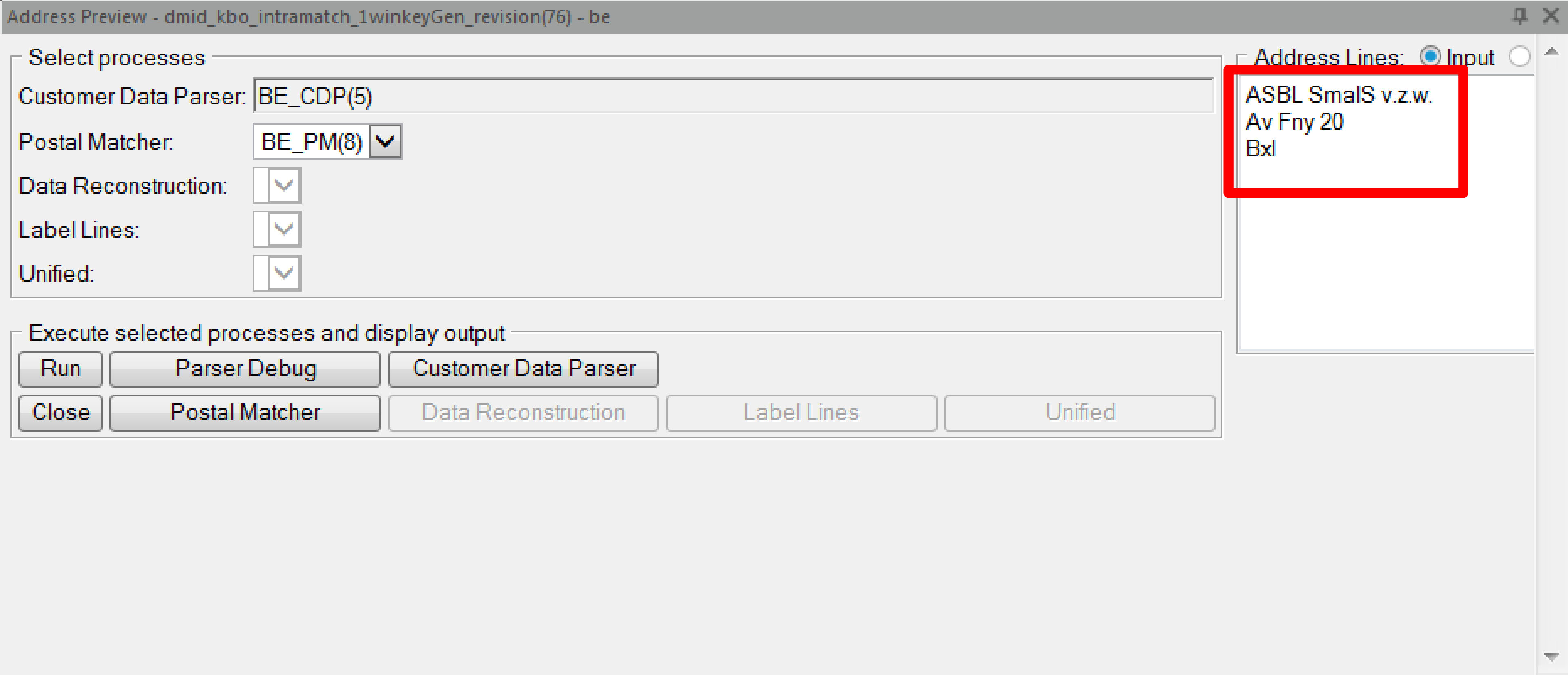
ASBL SmalS v.z.w.
Av Fny 20
Bxl

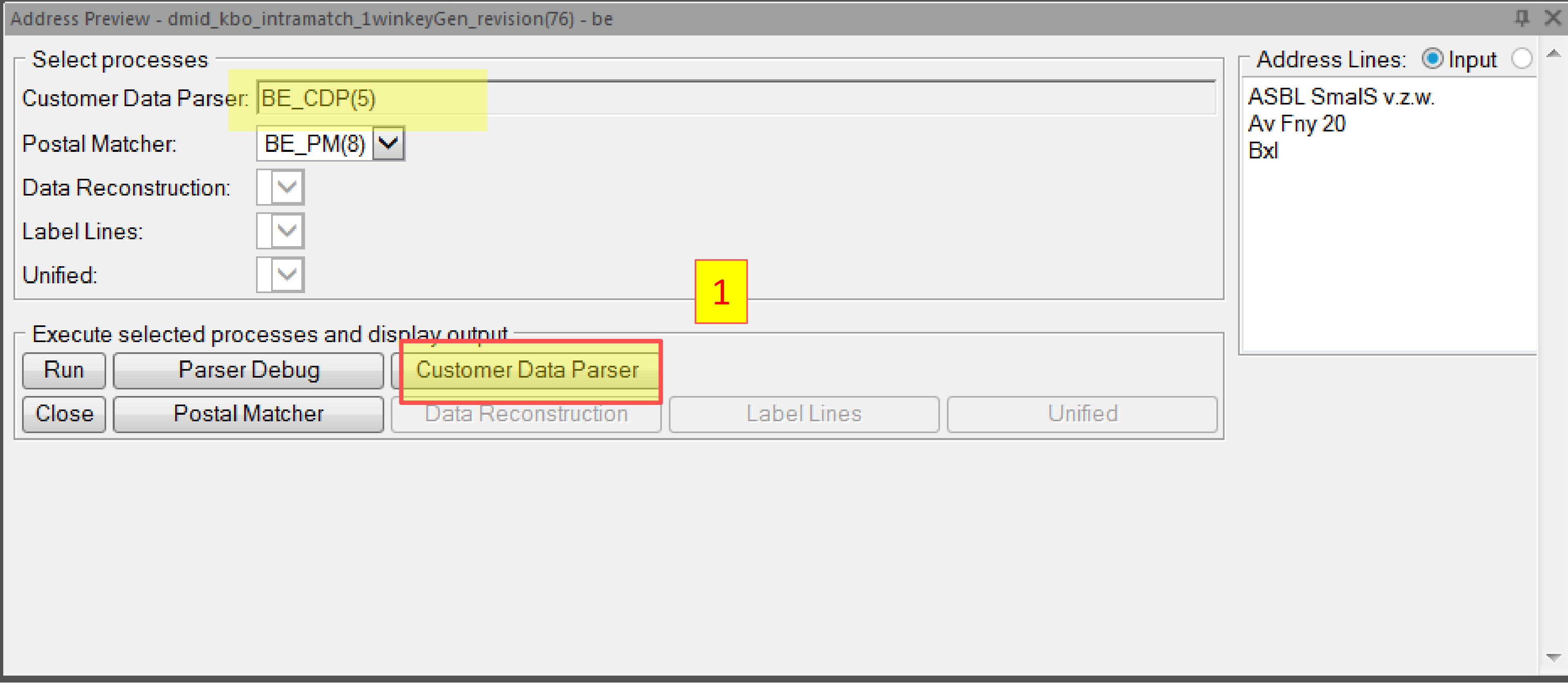
3.3. Parsing and postal validation hand in hand Let's try it !

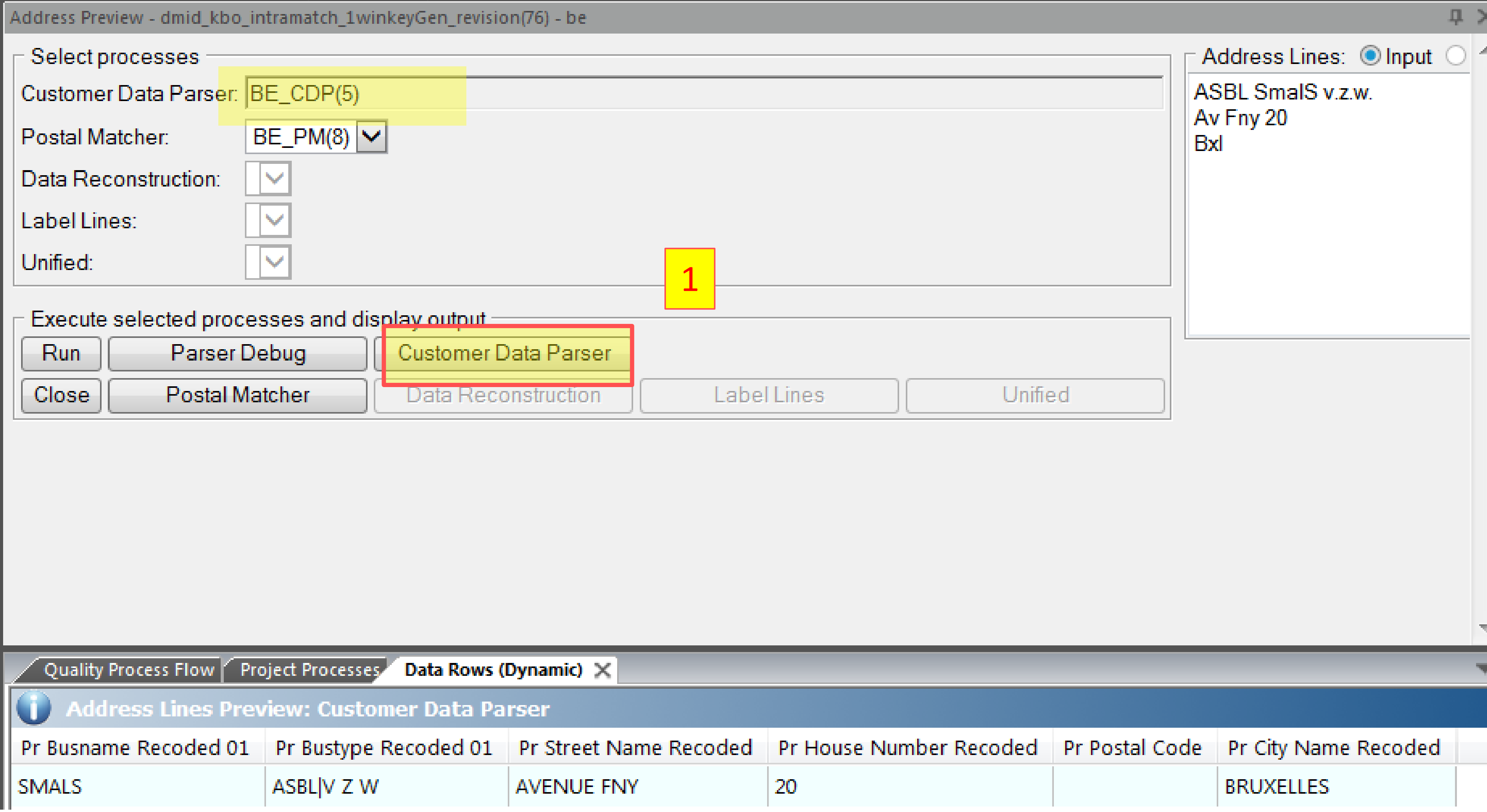
ASBL SmalS v.z.w.
Av Fny 20
Bxl

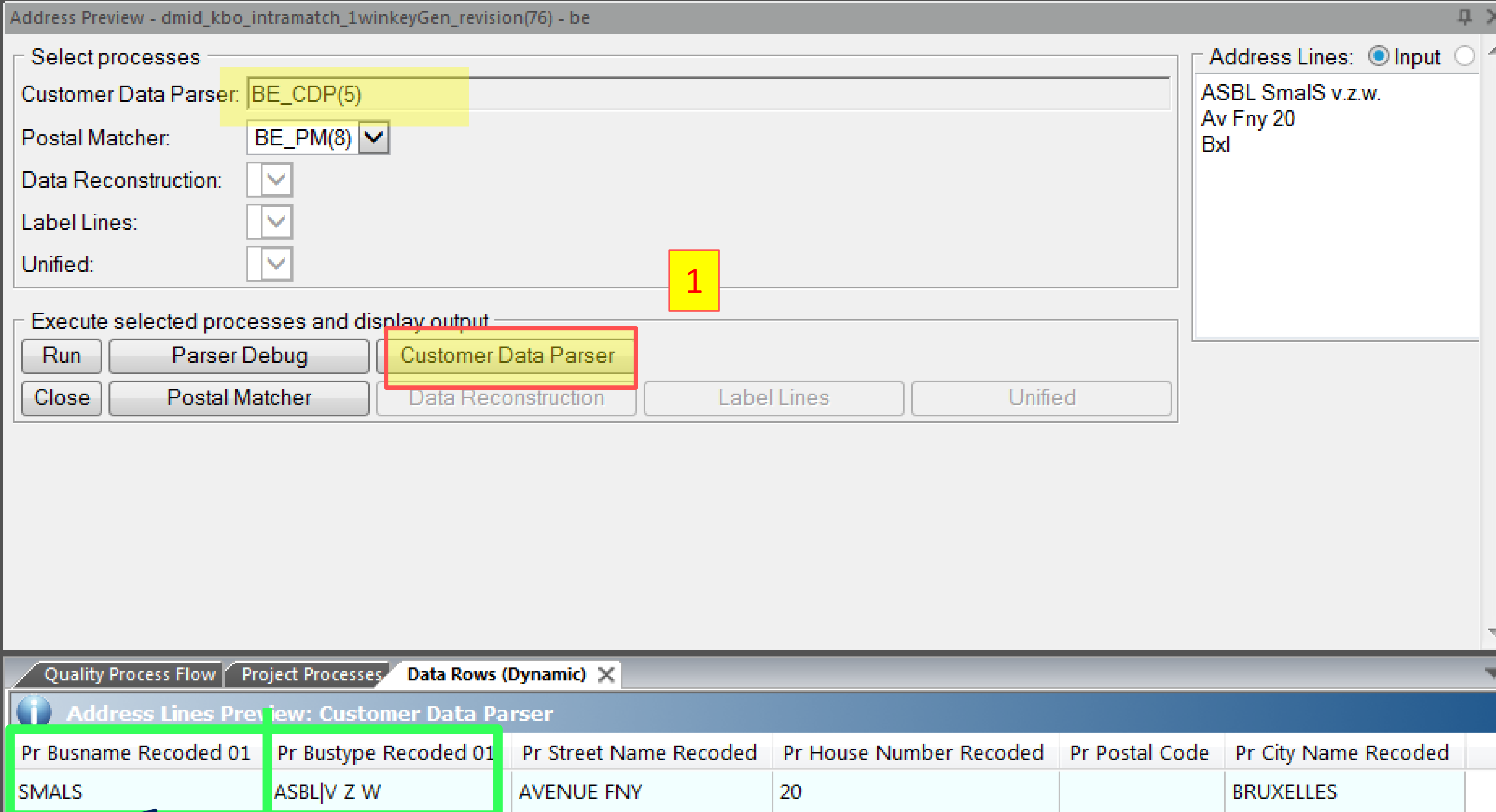
Unstandardized
-Denomination is inconsistent
-Wrong street
-No postcode
-City abbreviation


Icon Tools
Trillium Control Center - [Server: Public, Repository: dmining_id_first_tests, User: ganha]
Home
Advanced
Icon View
Style
Quality Process Flow
Project = dmid_kbo_intramatch_1winkeyGen_revision
FILTER_BE_A...
be tranfrmr
p55(3029)
:forpm
dress ...
BE_PM(8)
pmatch
Address ...
UNION_AN...

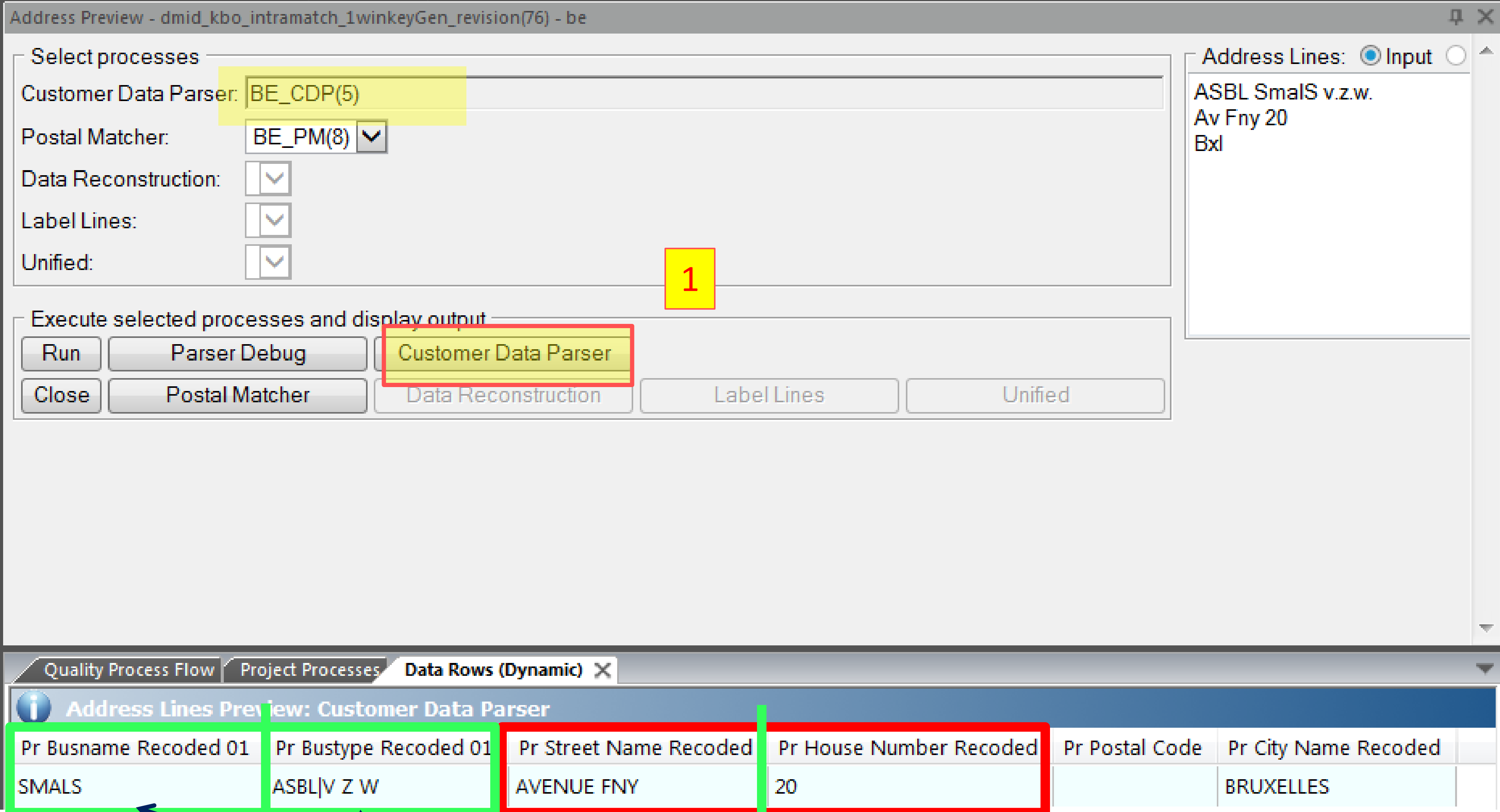

Address Preview - dmid_kbo_intramatch_1winkeyGen_revision(76) - be
Select processes
Customer Data Parser: BE_CDP(5)
Postal Matcher: BE_PM(8)
Data Reconstruction:
Label Lines:
Unified:
Execute selected processes and display output
Run
Parser Debug
Customer Data Parser
Close
Postal Matcher
Data Reconstruction
Label Lines
Unified
Address Lines: Input
ASBL SmalS v.z.w.
Av Fny 20
Bxl


Address Preview - dmid_kbo_intramatch_1winkeyGen_revision(76) - be
Select processes
Customer Data Parser: BE_CDP(5)
Postal Matcher: BE_PM(8)
Data Reconstruction:
Label Lines:
Unified:
1
Execute selected processes and display output
Run
Parser Debug
Customer Data Parser
Close
Postal Matcher
Data Reconstruction
Label Lines
Unified
Address Lines: Input
ASBL SmalS v.z.w.
Av Fny 20
Bxl


Address Preview - dmid_kbo_intramatch_1winkeyGen_revision(76) - be
Select processes
Customer Data Parser: BE_CDP(5)
Postal Matcher: BE_PM(8)
Data Reconstruction:
Label Lines:
Unified:
1
Address Lines: Input
ASBL SmalS v.z.w.
Av Fny 20
Bxl
Execute selected processes and display output
Run
Parser Debug
Customer Data Parser
Close
Postal Matcher
Data Reconstruction
Label Lines
Unified
Quality Process Flow
Project Processes
Data Rows (Dynamic)
Address Lines Preview: Customer Data Parser
Pr Busname Recoded 01
Pr Bustype Recoded 01
Pr Street Name Recoded
Pr House Number Recoded
Pr Postal Code
Pr City Name Recoded
SMALS
ASBL|V Z W
AVENUE FNY
20
BRUXELLES



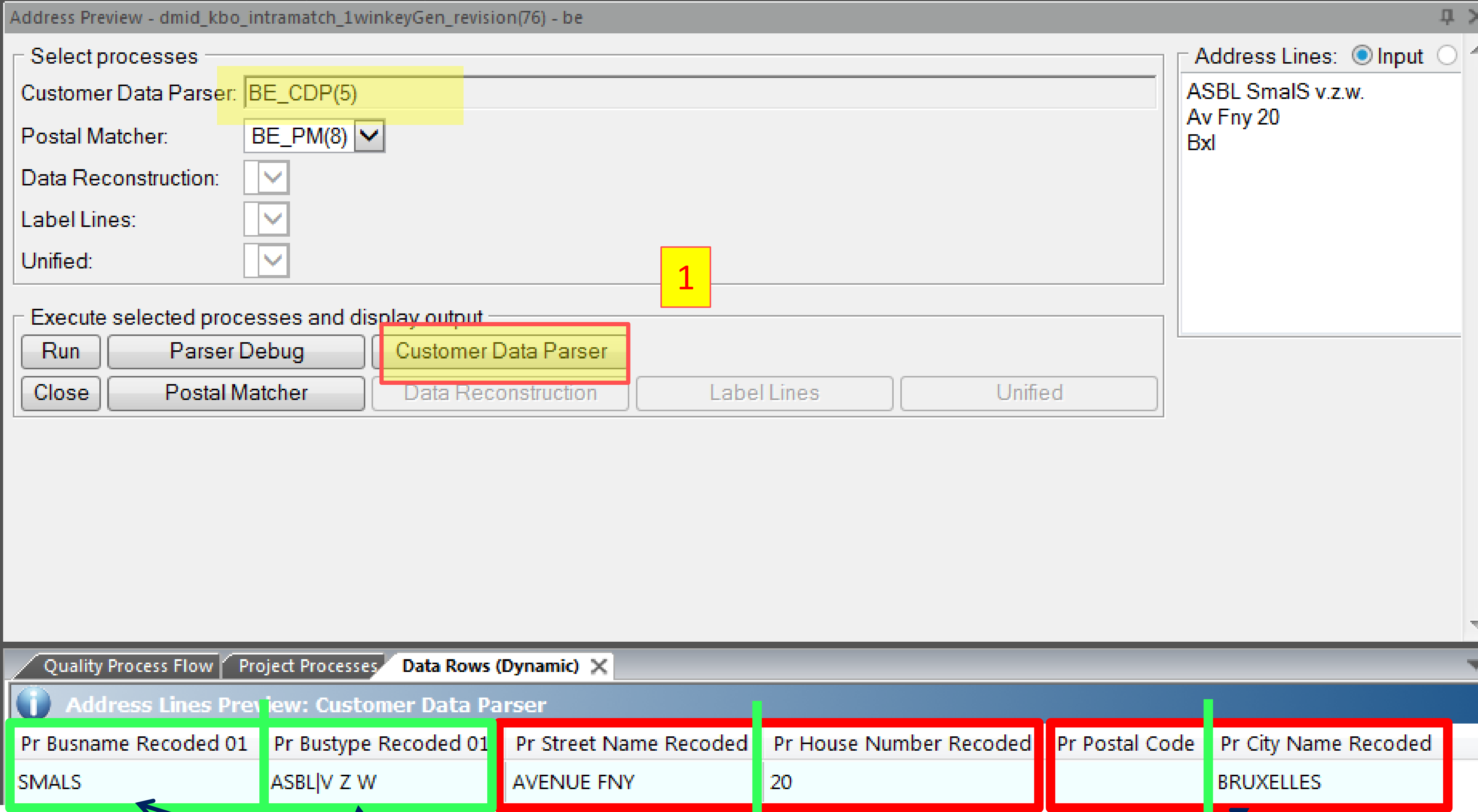
Name standardized, moving legal forms



Name standardized, moving legal forms

Input parsed into multiple attributes and standardized to upper case

ICT for society

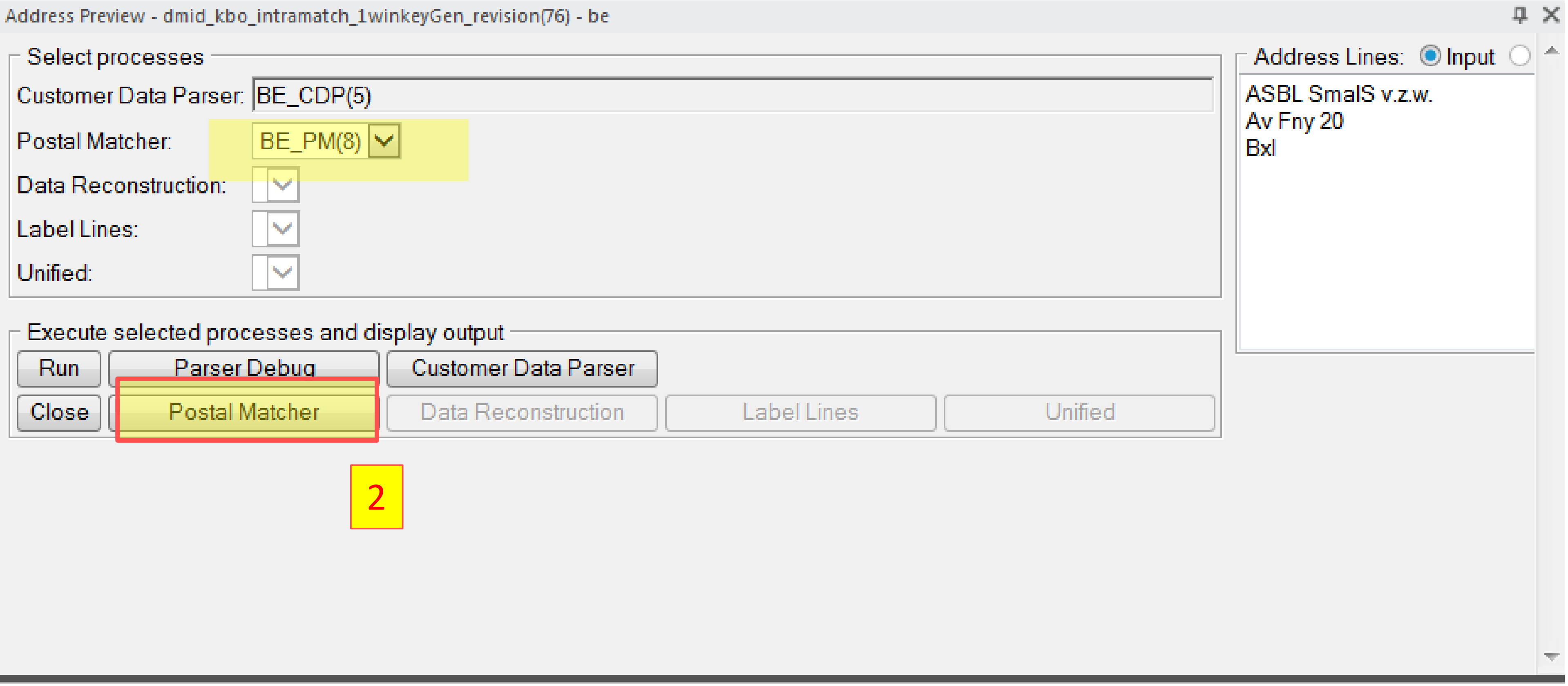


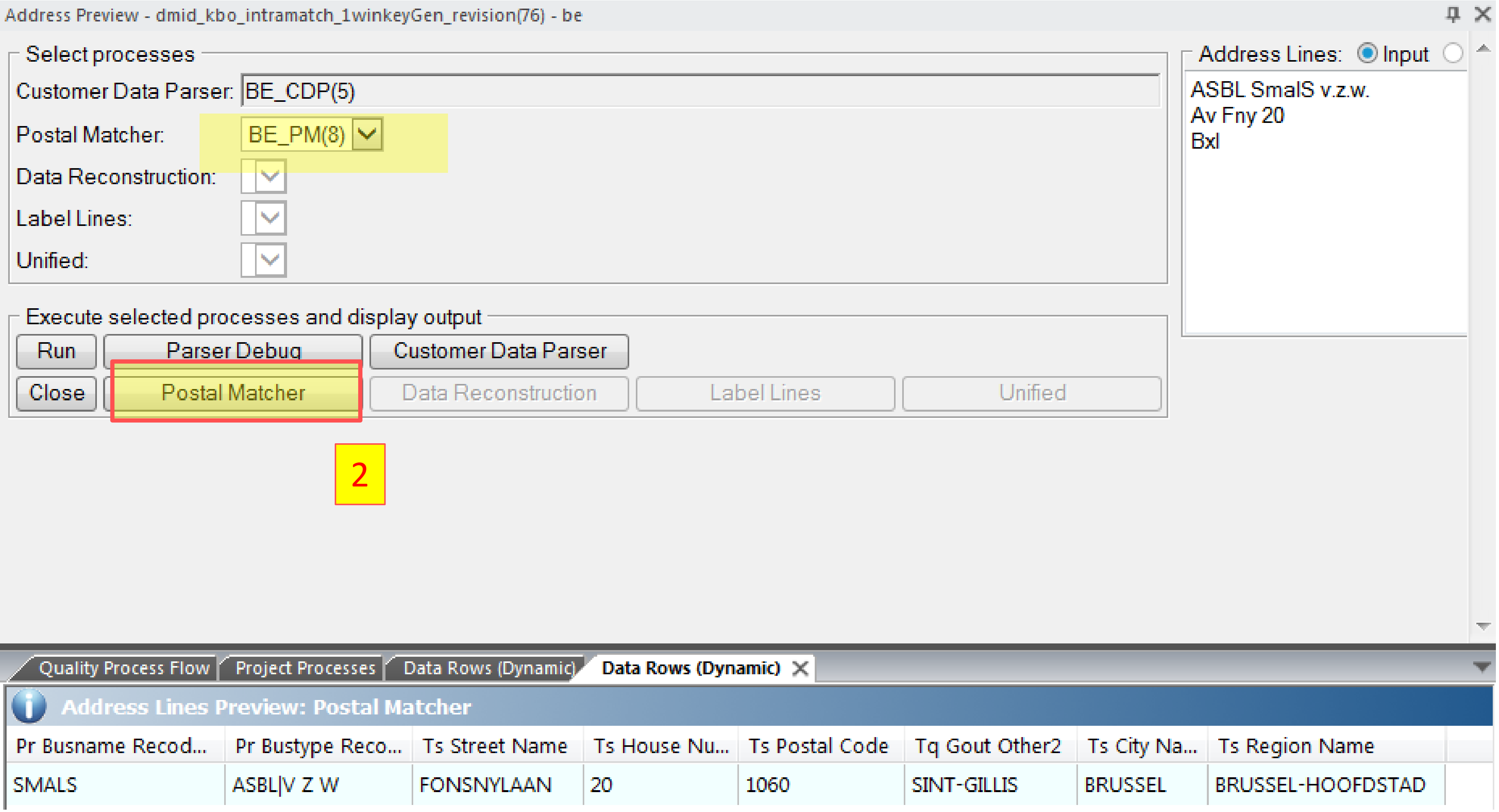
Name standardized, moving legal forms

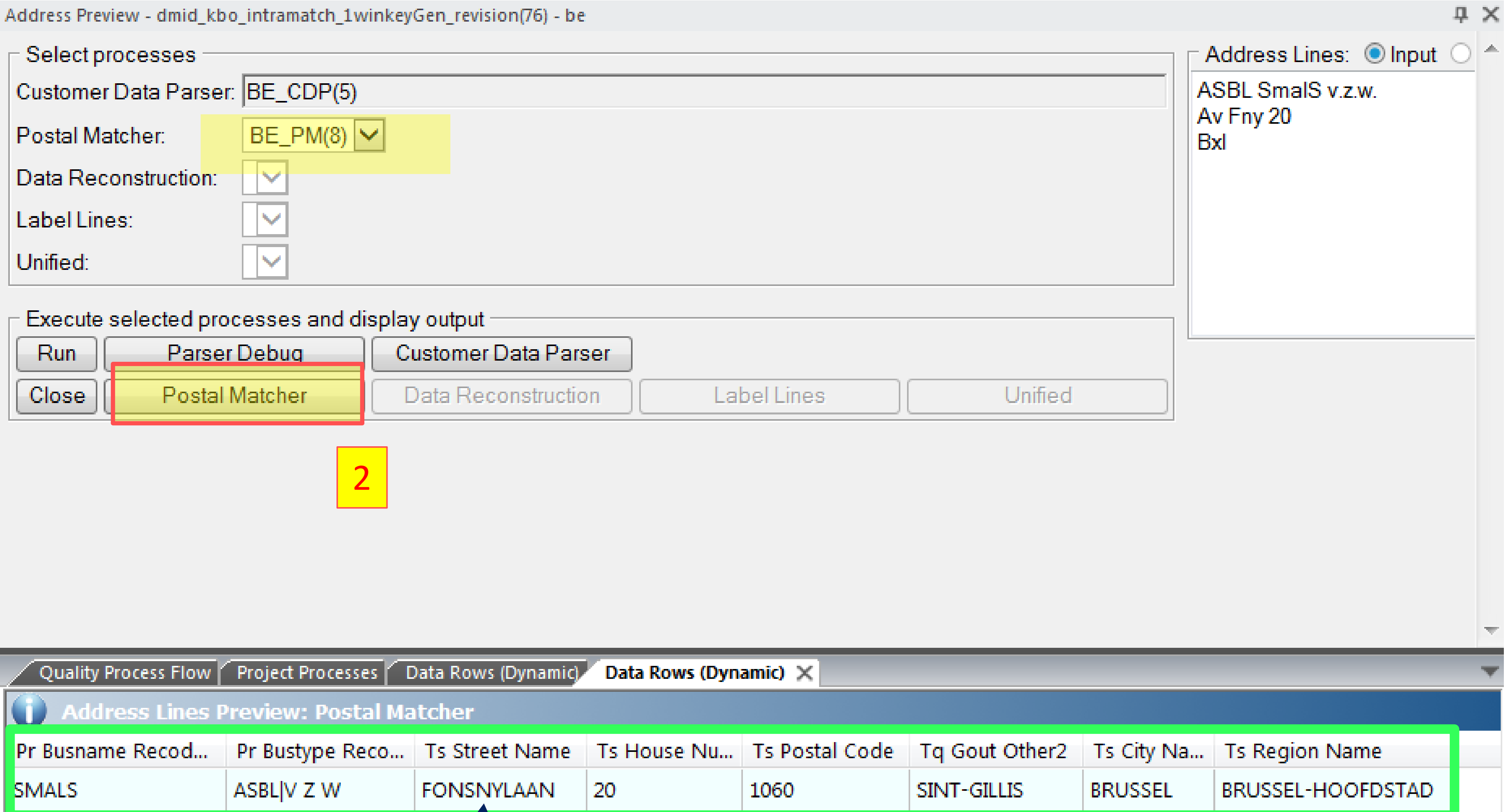
Input parsed into multiple attributes and standardized to upper case

City « Bxl » recoded to « BRUXELLES »

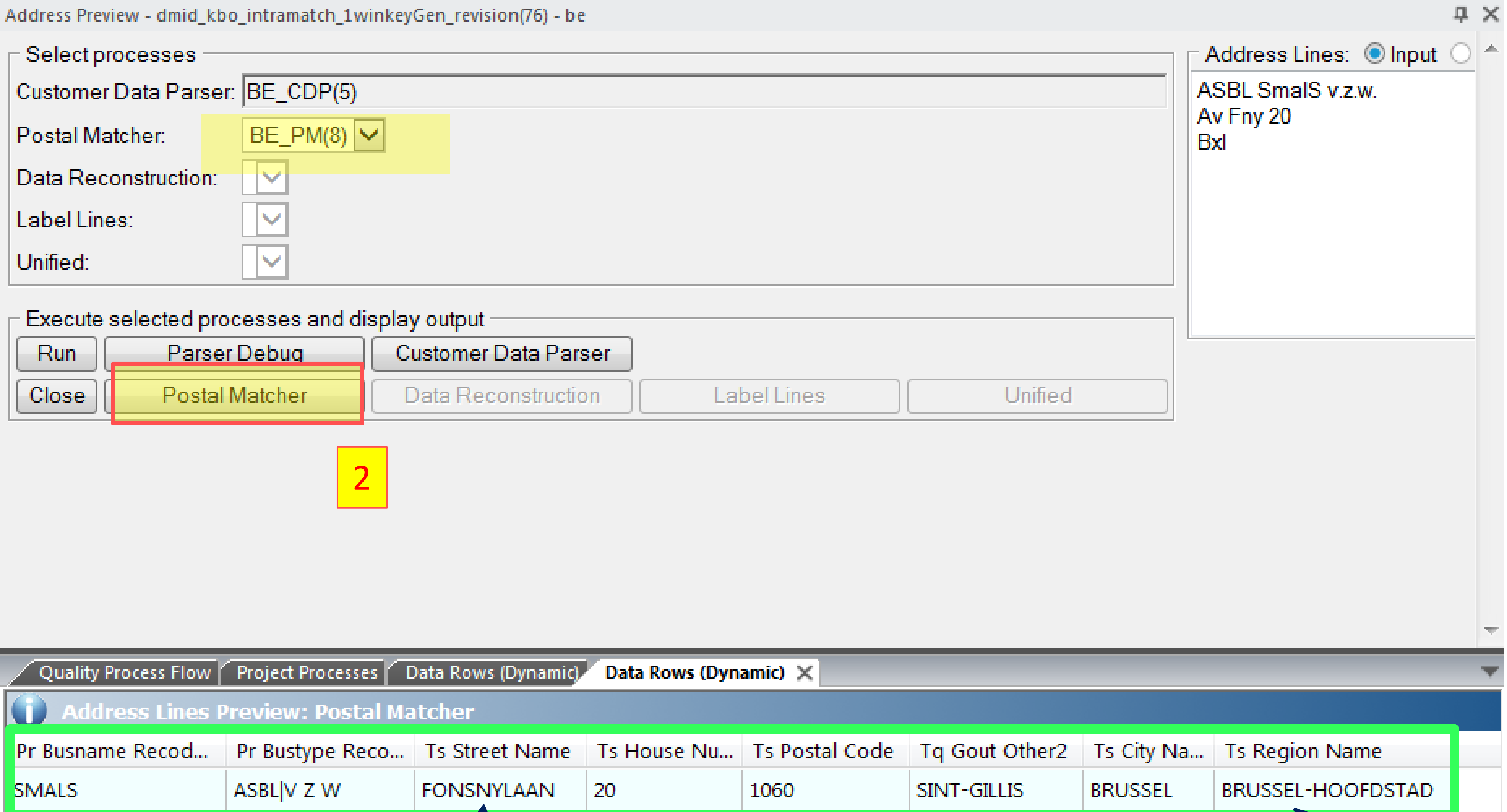
ICT for society


Address Preview - dmid_kbo_intramatch_1winkeyGen_revision(76) - be
Select processes
Customer Data Parser:
BE_CDP(5)
Postal Matcher:
BE_PM(8)
Data Reconstruction:
Label Lines:
Unified:
Execute selected processes and display output
Run
Parser Debug
Customer Data Parser
Close
Postal Matcher
Data Reconstruction
Label Lines
Unified
2
Address Lines:
Input
ASBL SmalS v.z.w.
Av Fny 20
Bxl


Address Preview - dmid_kbo_intramatch_1winkeyGen_revision(76) - be
Select processes
Customer Data Parser: BE_CDP(5)
Postal Matcher: BE_PM(8)
Data Reconstruction:
Label Lines:
Unified:
Address Lines: Input
ASBL SmalS v.z.w.
Av Fny 20
Bxl
Execute selected processes and display output
Run
Parser Debug
Customer Data Parser
Close
Postal Matcher
Data Reconstruction
Label Lines
Unified
2
Quality Process Flow
Project Processes
Data Rows (Dynamic)
Data Rows (Dynamic)
Address Lines Preview: Postal Matcher
Pr Busname Recod...
Pr Bustype Reco...
Ts Street Name
Ts House Nu...
Ts Postal Code
Tq Gout Other2
Ts City Na...
Ts Region Name
SMALS
ASBL|V Z W
FONSNYLAAN
20
1060
SINT-GILLIS
BRUSSEL
BRUSSEL-HOOFDSTAD

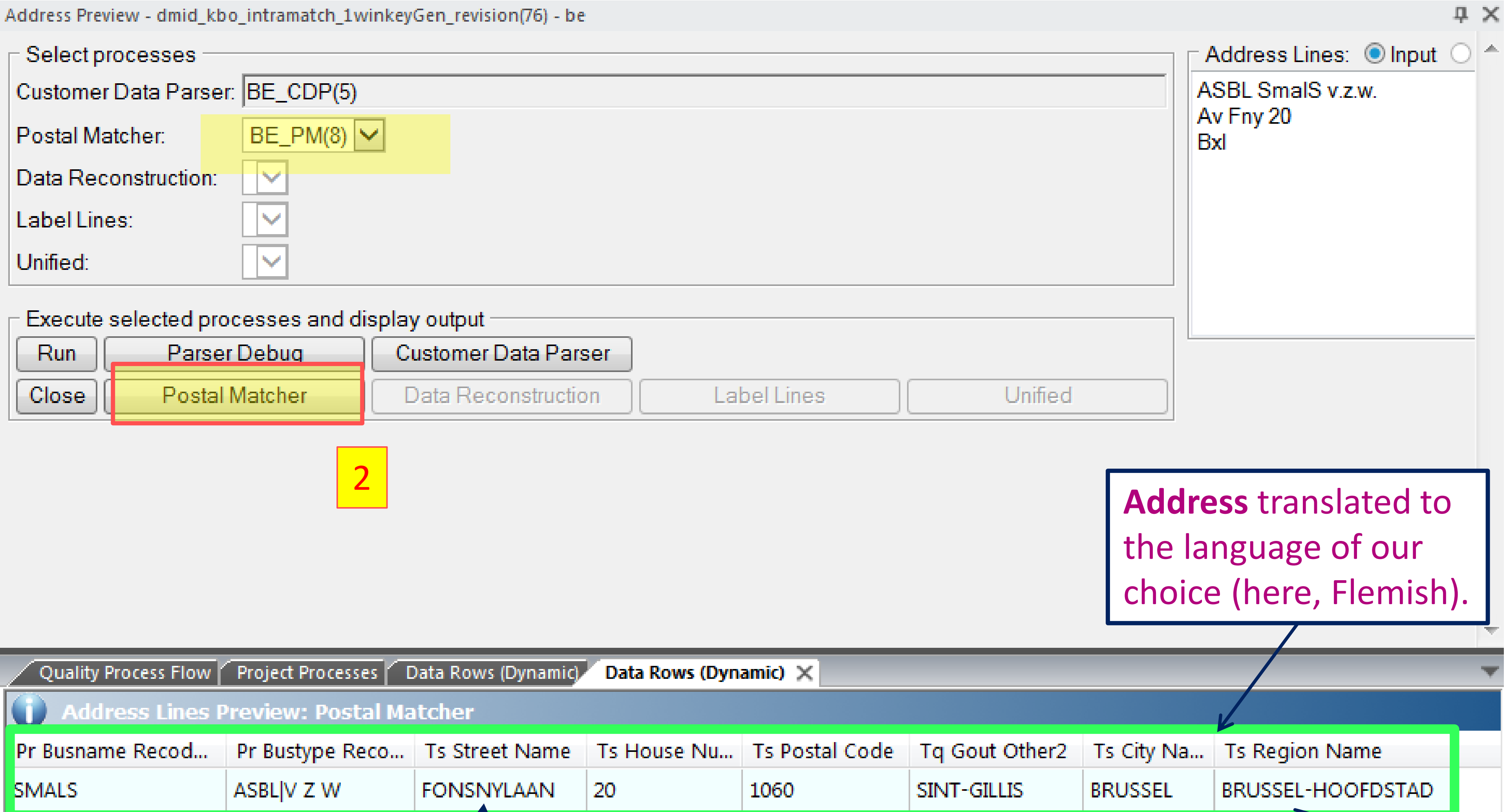


Street has been corrected



Street has been corrected

Postcode, municipality and region have been added (enrichment).



Address translated to the language of our choice (here, Flemish).

Street has been corrected

Postcode, municipality and region have been added (enrichment).

ULB MAsTIC Smals ICT for society

Extra: A problem to solve

Extra: A problem to solve

- Running a data flow, we noticed a big issue in the postal validation of our data

```
Record Failures
Count    Description
23506    Records Failed At State/City Level.
496712   Records Failed At Street Name Level.
4332     Records Failed At The House Number Level.
401      Records Failed At The Aggregate Components Level.
0        Records Failed Even Though They Matched, Because C
0        Records Passed But Directory Had Partial Or No Str

Postal Directory Date: JAN-2018
Address Accuracy Match: 4.6%
```

ULB MASTIC en

Smals ICT for society

Extra: A problem to solve

- Running a data flow, we noticed a big issue in the postal validation of our data

```
Record Failures
Count     Description
23506     Records Failed At State/City Level.
496712    Records Failed At Street Name Level.
4332      Records Failed At The House Number Level.
401       Records Failed At The Aggregate Components Level.
0         Records Failed Even Though They Matched, Because C
0         Records Passed But Directory Had Partial Or No Str

Postal Directory Date: JAN-2018
Address Accuracy Match: 4.6%
```

Extra: A problem to solve

- Drilling down into the data, we see:

Tsq Address
RUE DE CHIEVRES(T) 17
PLACE SAINTE-ANNE(COM) 21
PLACE COMMUNALE(L.L) S/N
PLACE COMMUNALE(L.L) 1
GRAND'PLACE(R) 1
GRAND'PLACE(L) 12
GRAND PLACE(BT) 11
RUE SAINT-PAUL(BIN) 14
GRAND'PLACE(CH) 13
PLACE ALBERT 1ER(FRO) 38
PLACE ALBERT 1ER(FRO) 38
GRAND-PLACE (MGS) 1
RUE SAINT MARTIN(MLC) 71

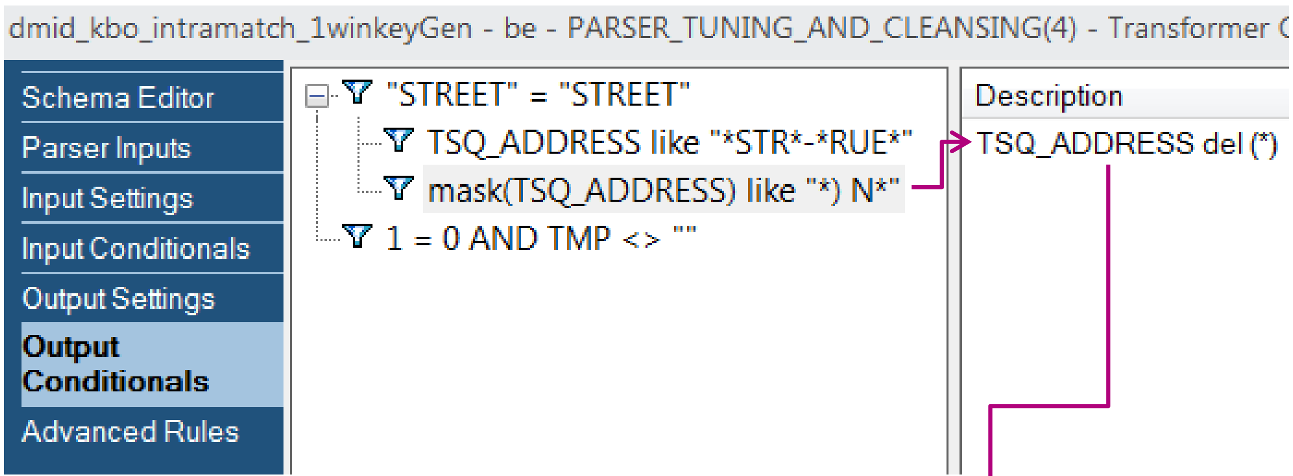
Extra: A problem to solve

- Drilling down into the data, we see:

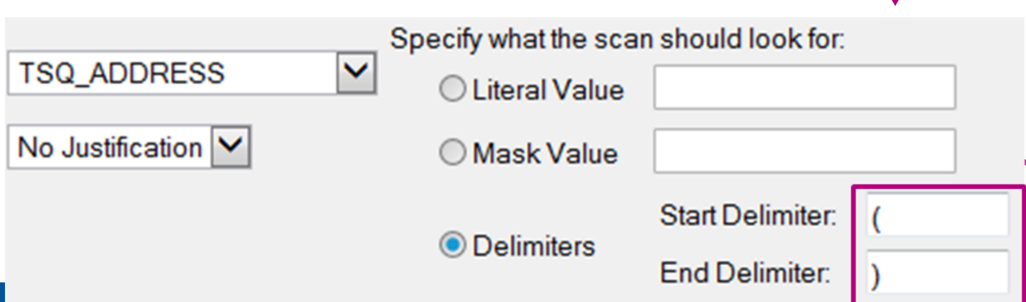
Tsq Address
RUE DE CHIEVRES(T) 17
PLACE SAINTE-ANNE(COM) 21
PLACE COMMUNALE(L.L) S/N
PLACE COMMUNALE(L.L) 1
GRAND'PLACE(R) 1
GRAND'PLACE(L) 12
GRAND PLACE(BT) 11
RUE SAINT-PAUL(BIN) 14
GRAND'PLACE(CH) 13
PLACE ALBERT 1ER(FRO) 38
PLACE ALBERT 1ER(FRO) 38
GRAND-PLACE (MGS) 1
RUE SAINT MARTIN(MLC) 71

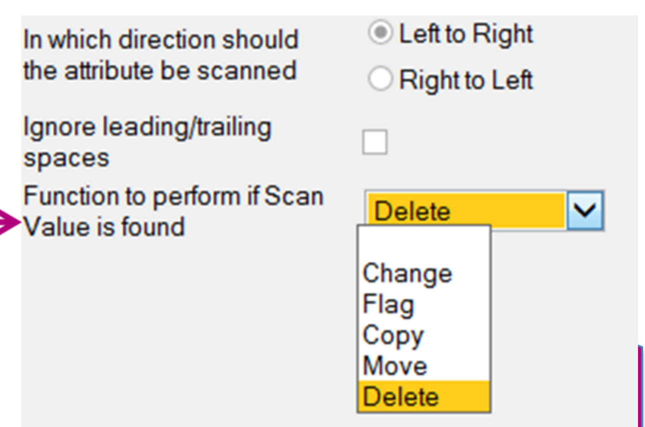
> 500 000 occurrences !!

Extra: A problem to solve



Conditional operation





ULB MA STIC
Smals ICT for society

Extra: A problem solved – Address correction : 4% → 96%)

- After re-running the flow starting with the Transformer
  - No pattern issues anymore in the Parser
  - **Spectacular rise in address validation**

```
Record Failures
Count    Description
2129     Records Failed At State/City Level.
15727    Records Failed At Street Name Level.
1251     Records Failed At The House Number Level.
310      Records Failed At The Aggregate Components Level.
0        Records Failed Even Though They Matched, Because
0        Records Passed But Directory Had Partial Or No St

Postal Directory Date: JAN-2018
Address Accuracy Match: 96.5%
```

ULB MASTIC en

Smals ICT for society

Extra: A problem solved – Address correction : 4% → 96%)

- Prevalence → Structural issue?

- Investigation
  - Application?
  - Public servants?
  - Certain cities or villages?

Data Standardisation: in summary

781114-269.56 | Yves Bontemps | Rue Prince Royal 102 Bruxelles

Parsing

781114-269.56 | Yves | Bontemps | Rue Prince Royal | 102 | Bruxelles

Correction & Enrichment

781114-269.56 | Yves | Bontemps | Rue **du** Prince Royal | 102 | **Ixelles**

Male

Koninklijke Prinsstraat

Elsene

1050

- Industry-grade DQ Tools come with
  - 10’s of thousands of rules for Parsing
  - knowledge bases for address enrichment, often region-sensitive

Source: Y. Bontemps, D. Van Dromme

Data standardization : in summary

- Programmation
  - Parsing
  - Enrichissement
- "Encodage" de la connaissance du domaine

```
public class PersonStandardiser {

    protected List<String> decomposeNN(String nationalNumber){
        List<String> decomposition = new ArrayList<String>(5);
        assert(nationalNumber.length() == 14);
        decomposition.add(nationalNumber.substring(0,2));
        decomposition.add(nationalNumber.substring(2,4));
        decomposition.add(nationalNumber.substring(4,6));
        decomposition.add(nationalNumber.substring(7,10));
        decomposition.add(nationalNumber.substring(11,13));
        return decomposition;
    }

    protected void enrich(Person pers, List<String> decomposedNN){
        if (Integer.parseInt(decomposedNN.get(3)) % 2 == 1){
            pers.setGender(Sex.MALE);
        }
        else{
            pers.setGender(Sex.FEMALE);
        }
        pers.setBirthDate(        decomposedNN.get(2)
                          +"/"+ decomposedNN.get(1)
                          +"/"+ decomposedNN.get(0)
                );
    }
```


Source: Y. Bontemps, D. Van Dromme

Contents

Contents

Contents

1. Data matching: main concepts

1. What is data matching?

2. Attribute level and Record level

3. Deterministic vs Probabilistic data matching

1.1. What is data matching?

- Linking between records
  - within one data source (duplicates detection)
  - across multiple sources (reference matching, detection of duplicates and inconsistencies)
  - even with different data models (data integration)

- …and deduplicating if needed
  - “Golden record” picking or commonization

- Some use cases
  - Creating a new repertory from external sources
  - Fusion between administrations
  - Integration of IT systems and DBs
  - Statistical modeling / datamining mixing referential and transactional data
  - Etc.

- ~ Relationship linking, entity matching, record linkage, entity resolution

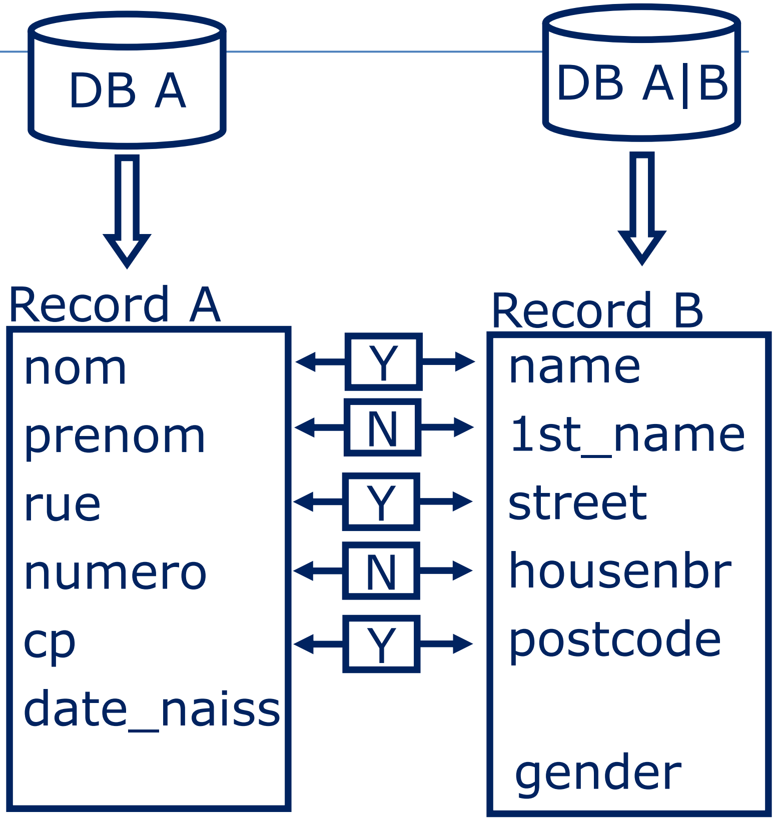
1.1. What is data matching?

- Dealing with fuzziness
  - typographical errors, inaccuracies, lack of standardization
  - != exact duplicates
- Agility critical task
  - definitions not clear from the start:
    what may or may not be considered as ‘double’ or ‘inconsistent’
  - many iterations with business are necessary
- Performance critical task
  - Support many iterations and application-critical deadlines
  - Esp. with millions of records
- Link with Anomaly Management
  - AM needs clear and formal definition (validated by business owners) of detection rules and treatment process

Source: Y. Bontemps, D. Van Dromme

1.2. Attribute level and Record level matching

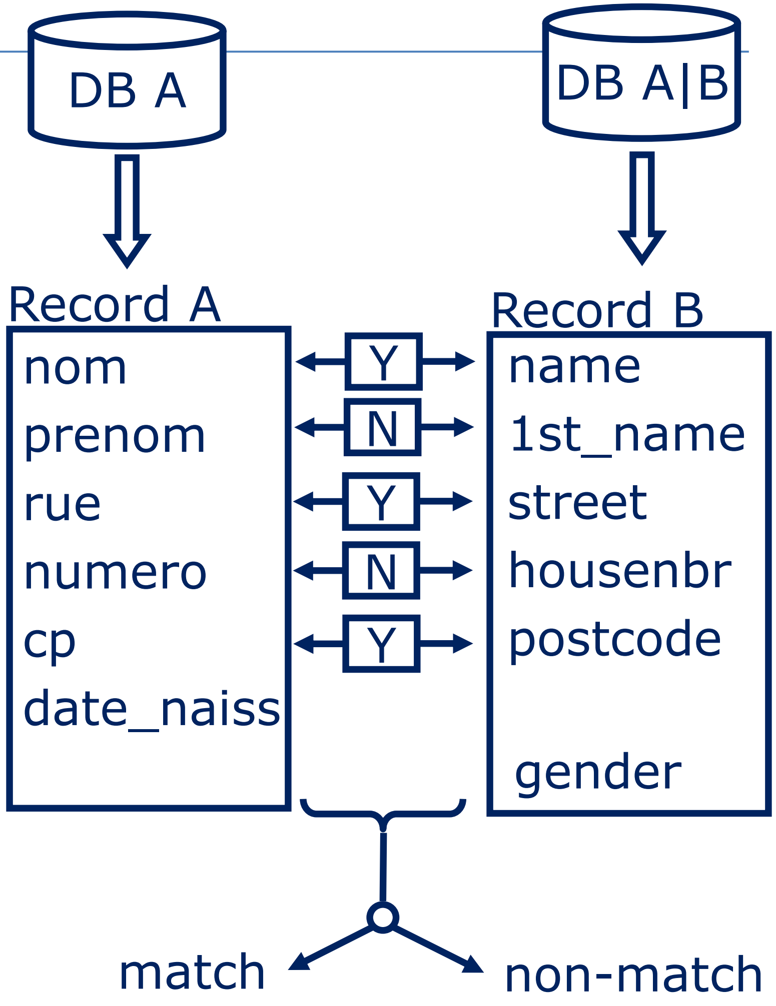
- Matching on two levels
  - Attribute-per-attribute: comparison algos



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

1.2. Attribute level and Record level matching

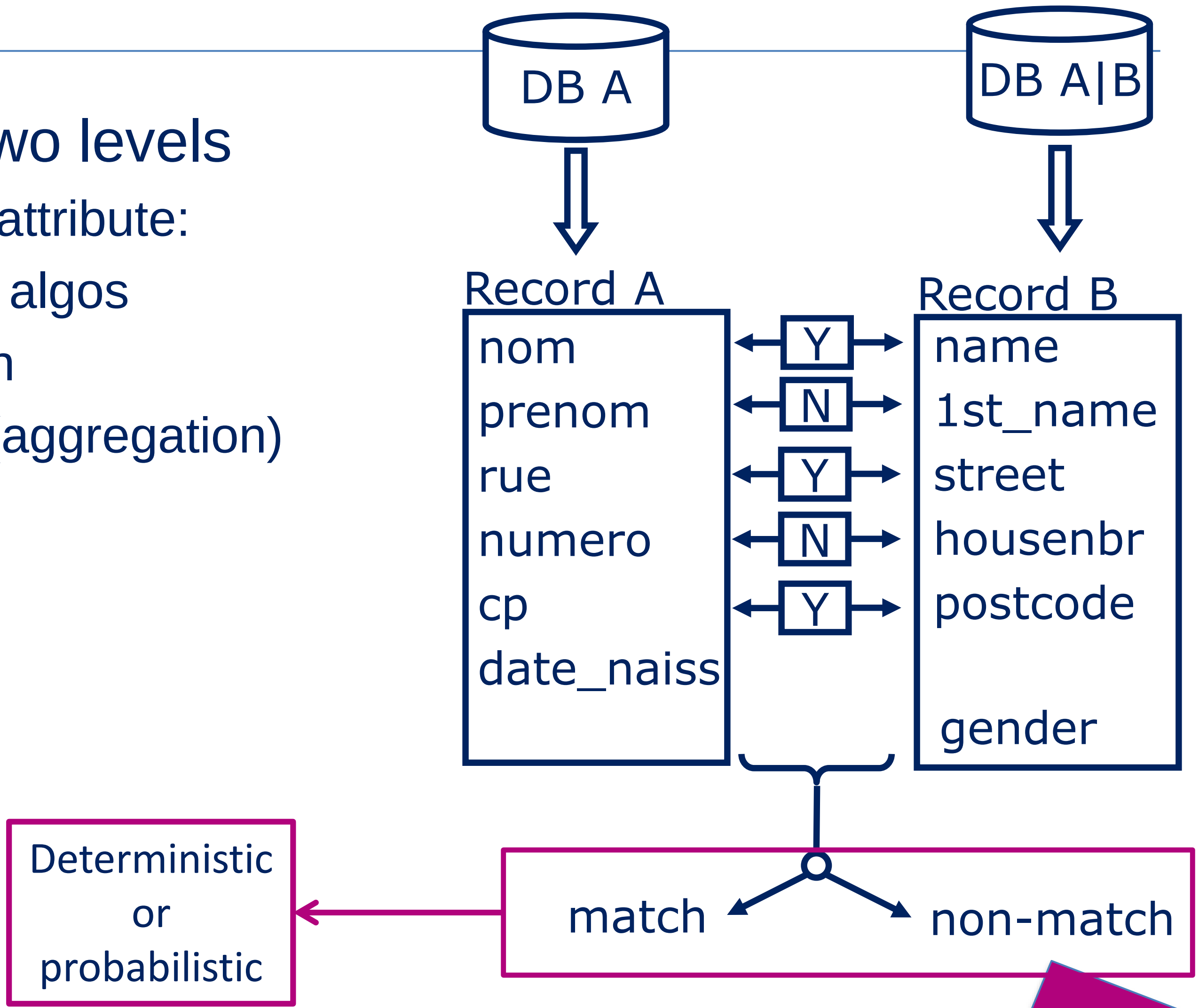
- Matching on two levels
  - Attribute-per-attribute:
    comparison algos
  - Then decision
    per record (aggregation)



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

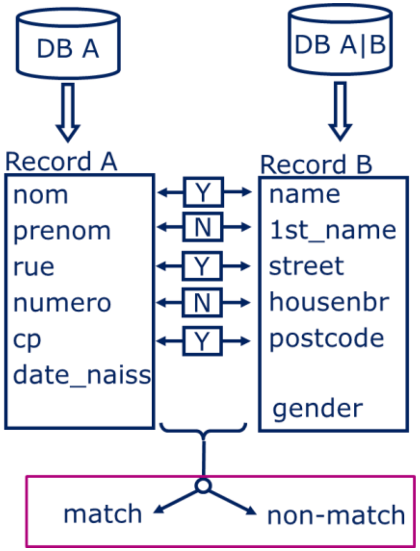
1.2. Attribute level and Record level matching

- Matching on two levels
  - Attribute-per-attribute:
    comparison algos
  - Then decision
    per record (aggregation)

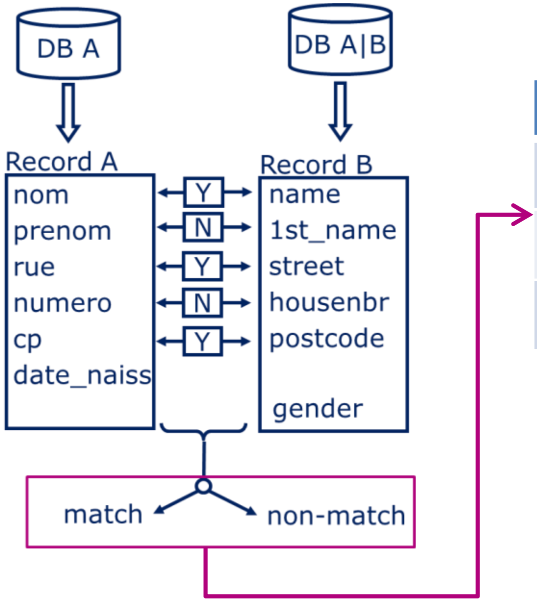


WINKLER W. E., *Overview of Record Linkage and Current Research Directions*, 2006.
http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

1.3. Deterministic vs Probabilistic matching



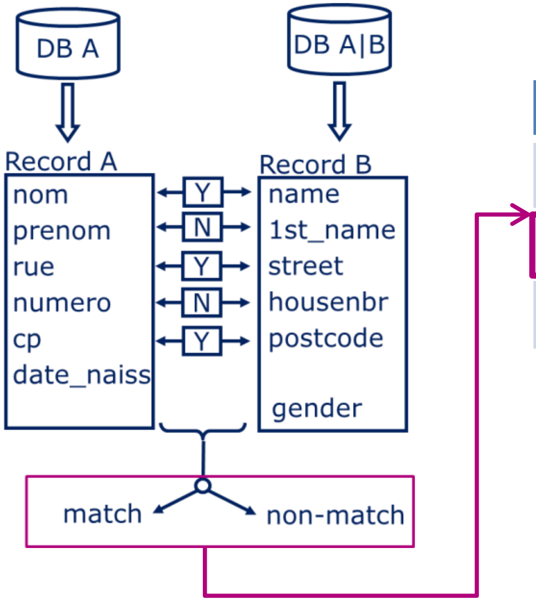
1.3. Deterministic vs Probabilistic matching



Deterministic: **match patterns** approach

Lastname	Firstname	Street	Housenb	Postcode	Decision
Y	Y	Y	-	Y	**Match**
Y	N	Y	N	Y	**Suspect**
Y	Y	N	N	-	**Fail**

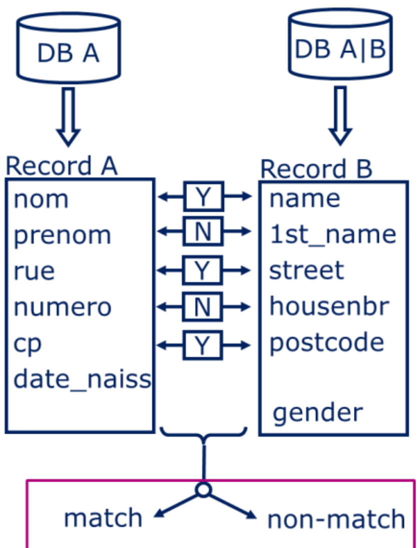
1.3. Deterministic vs Probabilistic matching



Deterministic: **match patterns** approach

Lastname	Firstname	Street	Housenb	Postcode	Decision
Y	Y	Y	-	Y	Match
Y	N	Y	N	Y	Suspect
Y	Y	N	N	-	Fail

1.3. Deterministic vs Probabilistic matching



Deterministic: **match patterns** approach

Lastname	Firstname	Street	Housenb	Postcode	Decision
Y	Y	Y	-	Y	**Match**
Y	N	Y	N	Y	**Suspect**
Y	Y	N	N	-	**Fail**

Probabilistic: **weighted attributes** approach (**!very simplified!** here)

Attribute	Weight
Lastname	0.40
Firstname	0.20
Street	0.35
Housenb	0.05
Postcode	0.10

~ 0.4 – 0.2 + 0.35 – 0.05 + 0.1 = 0.6

+

Threshold

match

0.8

0 0.45 1

Threshold

suspect

1.3. Probabilistic matching pros & cons

(+) Simplified human intervention (set weights, still w/ business)

(+) Native unmatch probability

(-) Can be difficult to understand or justify a match

→ Danger when dealing with e-gov data:

considerable impacts (human, legal, financial…)

(-) Weights and thresholds still imply a part of determinism (training or estimating)

1.3. Deterministic matching pros & cons

(+) Able to justify every step if legal requirements*

(+) Finer grain control and tuning

(-) Time needed for human iterations (business x IT)

(-) No native unmatch scoring

* “Deterministic” does not imply exact “==” matching
It simply means the decision (match vs non-match) is rule-based

2. Matching algorithms: families

/!\ blackbox software

Contents

2. Matching algorithms

- Any character strings
  - Names, streets, numbers, geographic coordinates… wherever there is fuzziness

- “Match”:
  - Does not mean exact match
  - Is entirely depending on the algorithm:
    - `Smals "=" Société de Mécanographie pour l’Application des Lois Sociales`
    - `Smals "=" Smals`
    - `Smals "=" SMALS`
    - `Smals "=" Smallz`
    - `Smals VZW "=" VZW Smals`

- Thousands of existing algorithms, always new ones
  - Generic or specific
  - Language agnostic or not
  - Called “comparison routines”, “clustering methods”, “matching functions”, etc.

- Valid in Deterministic AND Probabilistic approaches

2. Matching algorithms: families

Booleans / Classifiers	Rules & predicates	Phonetics
Similarity-based	Word-based	Token-based

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms: families

Booleans / Classifiers	**Rules & predicates**	**Phonetics**
Similarity-based	**Word-based**	**Token-based**

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms - Boolean family: Rules & predicates

- Booleans: they output Y or N (2 classes)
- Other classifiers: > 2 discrete output classes

There is no “in-between”

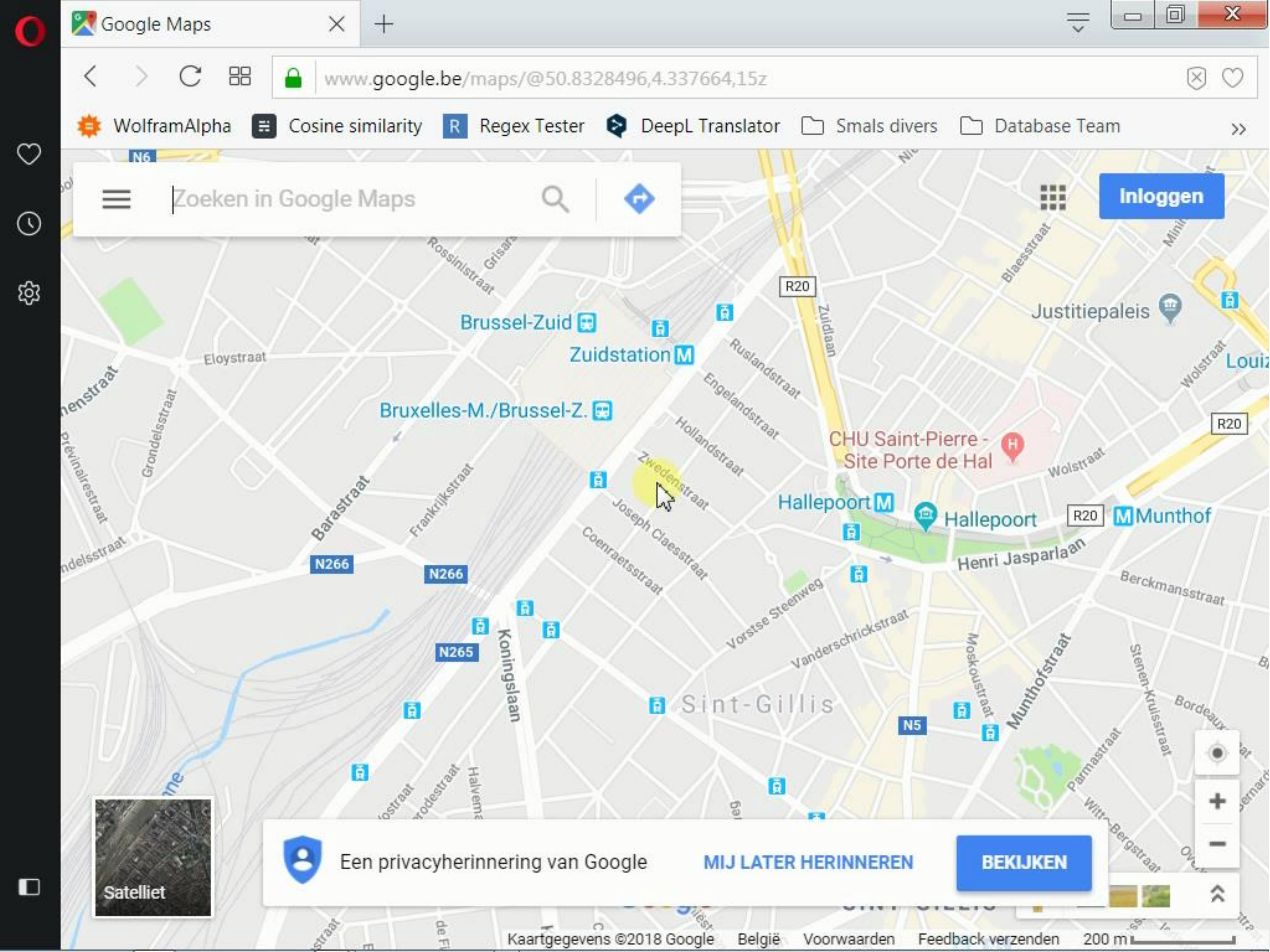
- Typically
  - Generic conventions (law, grammar, etc.)
  - Custom / domain-specific rules
  - Attribute B **is** <predicate> of Attribute A

- Boolean matching

if (rule(Attribute_A, Attribute_B))
then Attribute_A “=” Attribute_B

2. Matching algorithms - Boolean family: Rules & predicates (examples)

Attribute record A	Attribute record B	Algorithm	Output
Smals	Smals	Equal	Y
Translate	Translator	Stemming	Y
Bontemps	Bon	Prefix	Y
Bon**temps**	Temps	Suffix	Y
Vereniging **z**onder **w**instoogmerk	VZW	Initials	Y
…	…	…	…


Google Maps
www.google.be/maps/@50.8328496,4.337664,15z
WolframAlpha
Cosine similarity
Regex Tester
DeepL Translator
Smals divers
Database Team
Zoeken in Google Maps
Inloggen
Brussel-Zuid
Zuidstation
Bruxelles-M./Brussel-Z.
Justitiepaleis
CHU Saint-Pierre - Site Porte de Hal
Hallepoort
Hallepoort
Munthof
Henri Jasparlaan
Sint-Gillis
Eloystraat
Barastraat
Frankrijkstraat
Koningslaan
Zuidlaan
Ruslandstraat
Engelandstraat
Hollandstraat
Zwedenstraat
Joseph Claesstraat
Coenraetsstraat
Vorstse Steenweg
Vanderschrickstraat
Moskoustraat
Munthofstraat
Berckmansstraat
Blaesstraat
Wolstraat
Rossinistraat
Grondelsstraat
R20
N266
N265
N5
N6
Satelliet
Een privacyherinnering van Google
MIJ LATER HERINNEREN
BEKIJKEN
Kaartgegevens ©2018 Google
België
Voorwaarden
Feedback verzenden
200 m

2. Matching algorithms: families

Booleans / Classifiers	Rules & predicates	Phonetics
Similarity-based	Word-based	Token-based

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms - Boolean family: Phonetics

- Useful for
  - Transcription errors: oral → written
  - Typical pronuciation confusions (eg in Fr, “p”-“b”, “an”-“on”)
  - Post office counter, Call centers,…

- E.g. Mr “Dupont”:
  - Dupond
  - Dubont
  - Dubond
  - Dupant
  - …

- Phonetic matching:

  If (phon(Attribute_A) == phon(Attribute_B))
  then Attribute_A “=” Attribute_B

Rules & predicates | Phonetics | Word-based | Token-based

2. Matching algorithms - Boolean family: Phonetics (examples)

- Russel Soundex Algorithm (1918)

  1. Keep first character
  2. Delete a,e,h,i,o,u,w,y
  3. Recode:
     "1": B,F,P,V
     "2": C,G,J,K,Q,S,X
     "3": D,T
     "4": L
     "5": M, N
     "6": R
  4. Retain first 4 characters
     (padding with 0's if necessary)

Rules & predicates | Phonetics | Word-based | Token-based

Smals
ICT for society

2. Matching algorithms - Boolean family: Phonetics (examples)

- Russel Soundex Algorithm (1918)

  1. Keep first character
  2. Delete a,e,h,i,o,u,w,y
  3. Recode:
     "1": B,F,P,V
     "2": C,G,J,K,Q,S,X
     "3": D,T
     "4": L
     "5": M, N
     "6": R
  4. Retain first 4 characters
     (padding with 0's if necessary)

- Example

```
Dupont      Dubond
1. D        1. D
2. DPNT     2. DBND
3. D153     3. D153
```

Rules & predicates | Phonetics | Word-based | Token-based

2. Matching algorithms - Boolean family: Phonetics (examples)

- And many others:
  - Metaphone (no length limit)
  - Double Metaphone (language specificities)
  - NYSIIS (US English names)
  - Daitch-Mokotoff
    - Slavic & Yiddic languages
    - 54 entries
  - Fonem
    - French-oriented
    - 64 rules
  - Phonex
  - etc.

Smals
ICT for society

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms - Boolean family: Phonetics (examples)

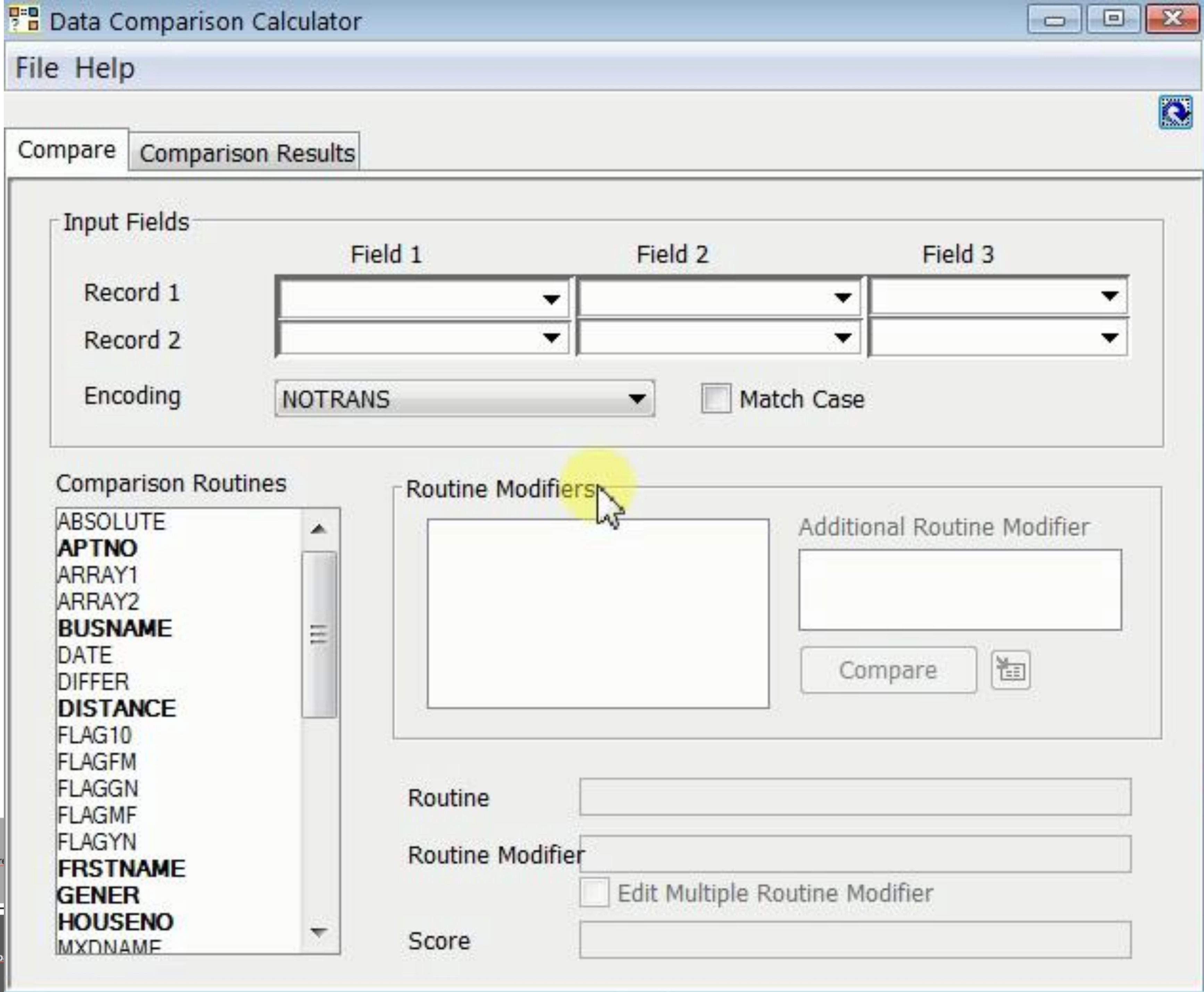
A	B	Algorithm	Algo(A)	Algo(B)	Output
Standard & Poor's	Standard de Liège	Soundex	S353	S353	Y
		Metaphone	STNTRTPRS	STNTRTTLJ	N

2. Matching algorithms - Boolean family: Phonetics (examples)

A	B	Algorithm	Algo(A)	Algo(B)	Output
Standard & Poor's	Standard de Liège	Soundex	S353	S353	Y
		Metaphone	STNTRTPRS	STNTRTTLJ	**N**

2. Matching algorithms - Boolean family: Phonetics (examples)

A	B	Algorithm	Algo(A)	Algo(B)	Output
Standard & Poor's	Standard de Liège	Soundex	S353	S353	Y
		Metaphone	STNTRTPRS	STNTRTTLJ	**N**
McBridge	MacBrigge	Metaphone	MKBRJ	MKBRK	N


Data Comparison Calculator
File Help
Compare
Comparison Results
Input Fields
Field 1
Field 2
Field 3
Record 1
Record 2
Encoding
NOTRANS
Match Case
Comparison Routines
ABSOLUTE
APTNO
ARRAY1
ARRAY2
BUSNAME
DATE
DIFFER
DISTANCE
FLAG10
FLAGFM
FLAGGN
FLAGMF
FLAGYN
FRSTNAME
GENER
HOUSENO
Routine Modifiers
Additional Routine Modifier
Compare
Routine
Routine Modifier
Edit Multiple Routine Modifier
Score

2. Matching algorithms - Boolean family: Phonetics (examples)

A	B	Algorithm	Algo(A)	Algo(B)	Output
Standard & Poor's	Standard de Liège	Soundex	S353	S353	Y
		Metaphone	STNTRTPRS	STNTRTTLJ	**N**
McBridge	MacBrigge	Metaphone	MKBRJ	MKBRK	N
		NYSIIS	MCBRAG	MCBRAG	**Y**
…	…	…	…	…	**…**

Rules & predicates | Phonetics

Word-based | Token-based

2. Matching algorithms: families

Booleans / Classifiers	Rules & predicates	Phonetics
Similarity-based	Word-based	Token-based

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms – Similarity family: Word-based

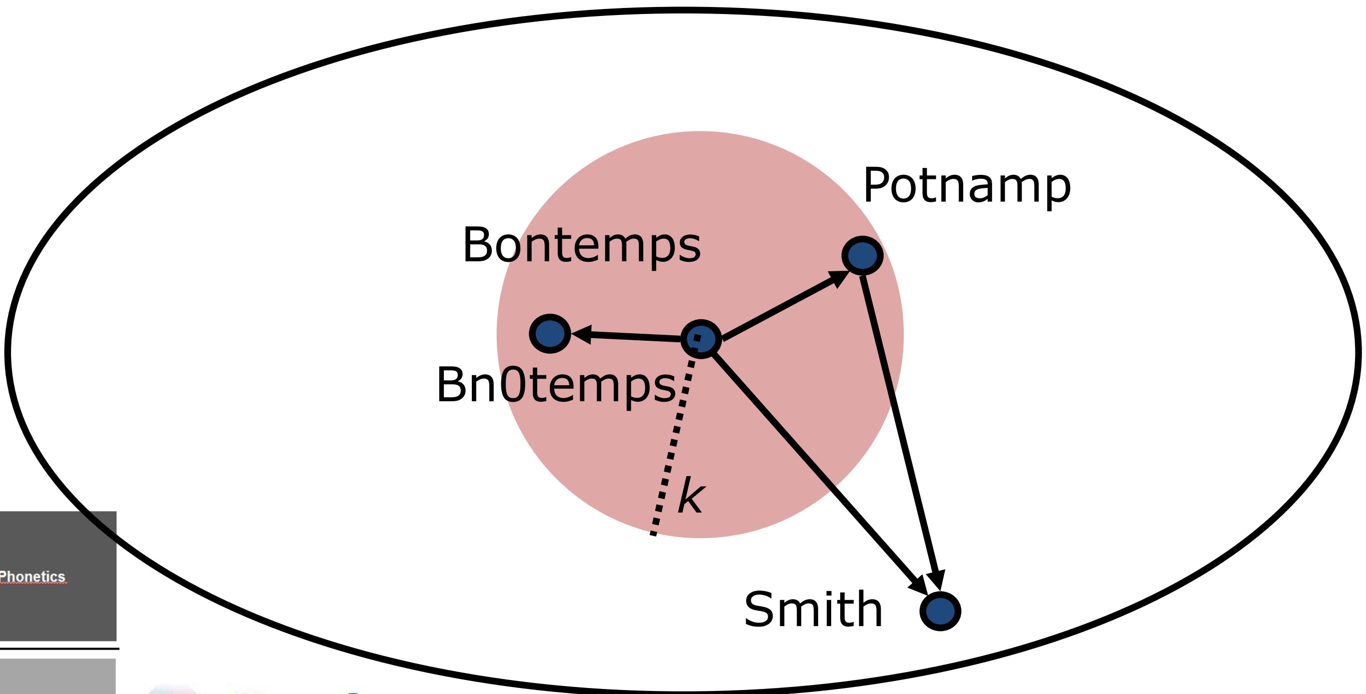
- Also called "distance algorithms"
  - Same principle
  - Inverted approach: instead of measuring similarity, you measure differences

- Useful for typos, errors in OCR, etc.

- Output: continuous score
  - More granularity than boolean algos
  - Integer / Float between -1 and 1, 0 and 1, 0 and 100, etc…
  - Implementation specific

2. Matching algorithms – Similarity family: Word-based

- Matching based on a distance algorithm:
  - if (distance(Attribute_A , Attribute_B) <= k)
    then Attribute_A “=” Attribute_B



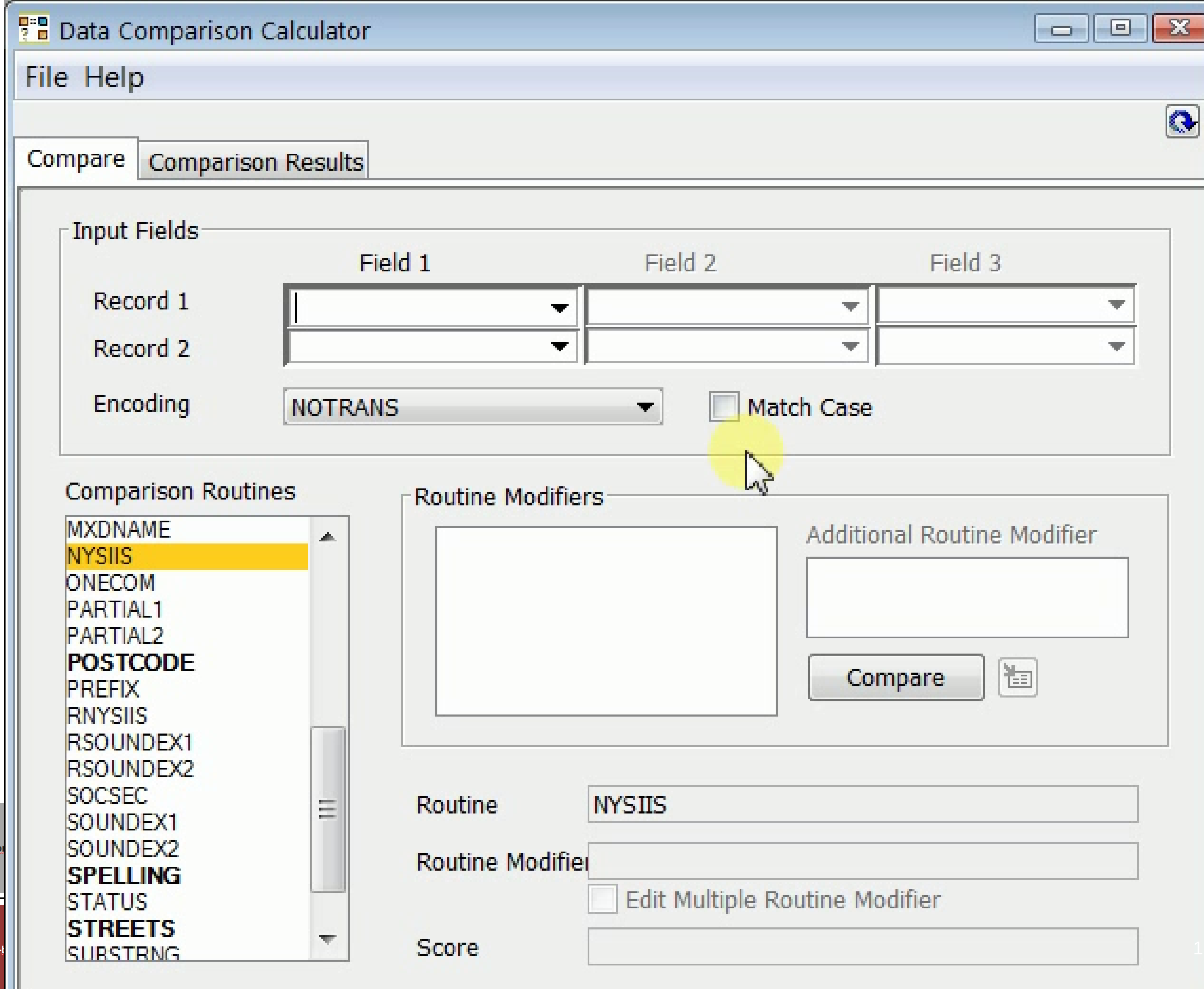
Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms – Similarity family: Word-based (examples)

- Levenshtein distance: min. number of « operations » to transform Attribute_B into Attribute_A
  - Insertion (I)
  - Deletion (D)
  - Substitution (S)
  - (Damereau-Levenshtein : Transposition (T))

- Example
  - Attribute_A = Smals, Attribute_B = Smallz
  - Smallz (D « l »)
  - Smals (S « z » → « s »)

→ 2 operations


Data Comparison Calculator
File Help
Compare
Comparison Results
Input Fields
Field 1
Field 2
Field 3
Record 1
Record 2
Encoding
NOTRANS
Match Case
Comparison Routines
MXDNAME
NYSIIS
ONECOM
PARTIAL1
PARTIAL2
POSTCODE
PREFIX
RNYSIIS
RSOUNDEX1
RSOUNDEX2
SOCSEC
SOUNDEX1
SOUNDEX2
SPELLING
STATUS
STREETS
Routine Modifiers
Additional Routine Modifier
Compare
Routine
NYSIIS
Routine Modifier
Edit Multiple Routine Modifier
Score

2. Matching algorithms: families

Booleans / Classifiers	Rules & predicates	Phonetics
Similarity-based	Word-based	Token-based

Source: Y. Bontemps, D. Van Dromme

2. Matching algorithms – Similarity family: Token-based

- Like word-based algos, output: continuous score

- Specific use of token-based approach:
  - Token = "atomic unit of language" (mostly, "words")
  - Comparing tokens != comparing whole string
  - Most often, word order does not count
  - Possibly take discriminative power of tokens into account
    - If rare token matches, weighs more than another token match (e.g. TF-IDF)

- Token-based matching

  If (token(Attribute_A, Attribute_B) >= thresh)

  then Attribute_A = Attribute_B

2. Matching algorithms – Similarity family: Token-based (examples)

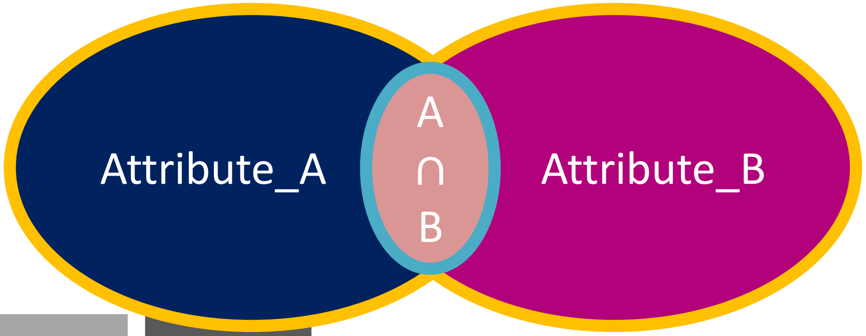
- Jaccard index
  - Given Attribute_A , Attribute_B, Jaccard(Attribute_A,Attribute_B):

$$\frac{|Attribute\_A \cap Attribute\_B|}{|Attribute\_A \cup Attribute\_B|}$$

2. Matching algorithms – Similarity family: Token-based (examples)

- Jaccard index
  - Given Attribute_A , Attribute_B, Jaccard(Attribute_A,Attribute_B):

$$\frac{|Attribute\_A \cap Attribute\_B|}{|Attribute\_A \cup Attribute\_B|}$$

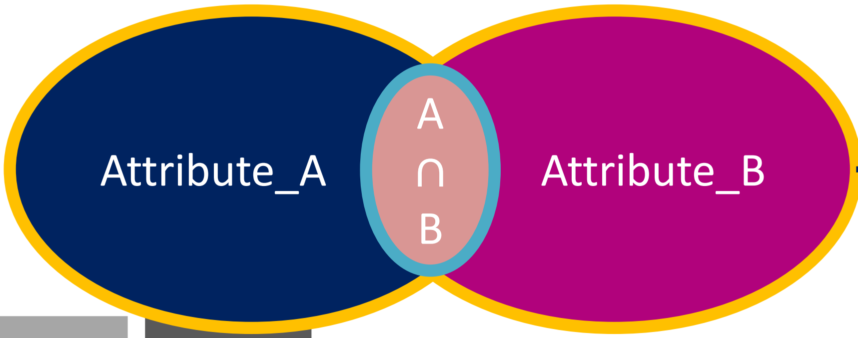


Rules & predicates | Phonetics
Word-based | Token-based

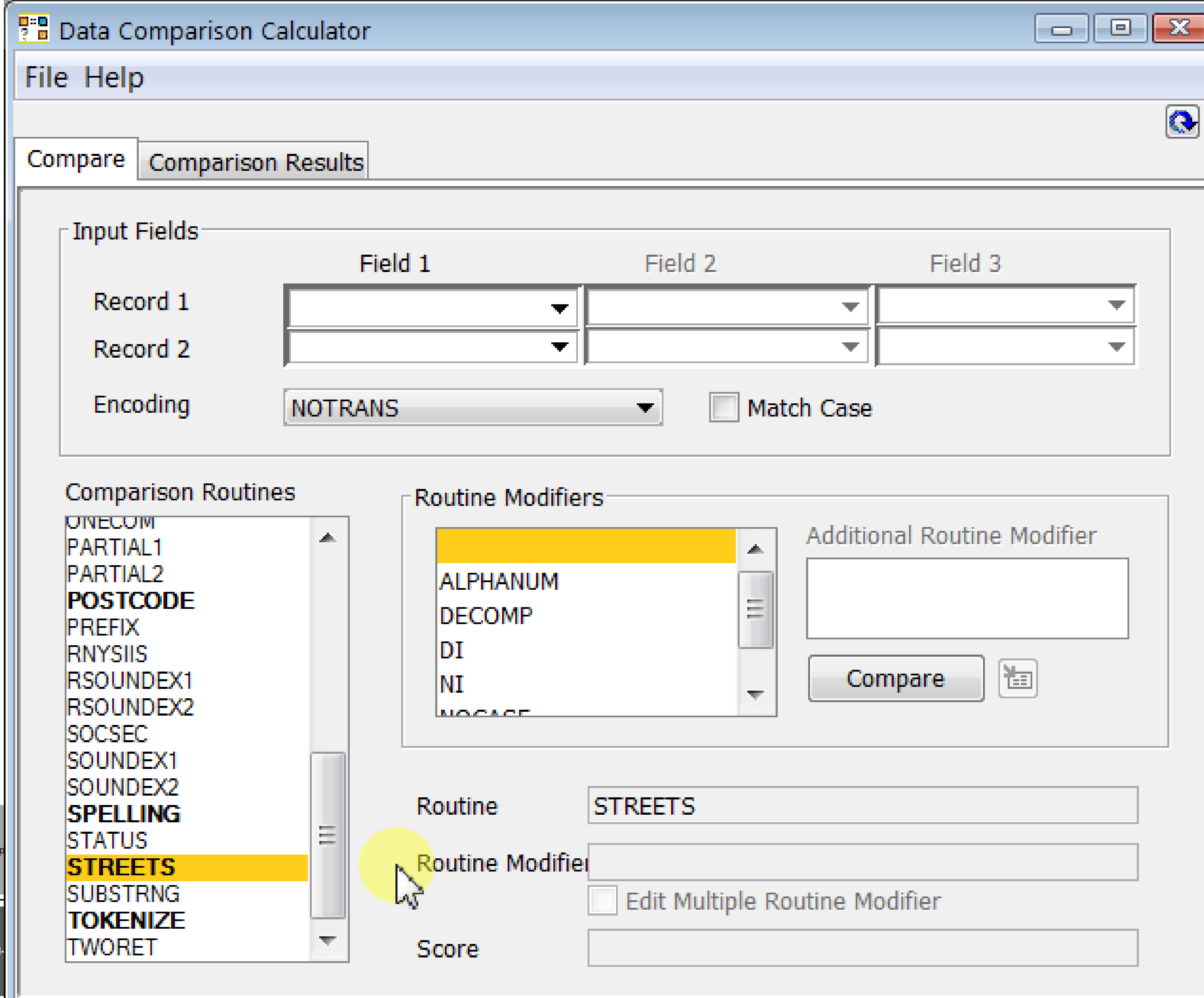
2. Matching algorithms – Similarity family: Token-based (examples)

- Jaccard index
  - Given Attribute_A , Attribute_B, Jaccard(Attribute_A,Attribute_B):

$$\frac{|Attribute\_A \cap Attribute\_B|}{|Attribute\_A \cup Attribute\_B|}$$



Attributes	Jaccard index
Smals VZW ASBL	$\frac{2}{3}$
Smals VZW	


Data Comparison Calculator
File Help
Compare
Comparison Results
Input Fields
Field 1
Field 2
Field 3
Record 1
Record 2
Encoding
NOTRANS
Match Case
Comparison Routines
PARTIAL1
PARTIAL2
POSTCODE
PREFIX
RNYSIIS
RSOUNDEX1
RSOUNDEX2
SOCSEC
SOUNDEX1
SOUNDEX2
SPELLING
STATUS
STREETS
SUBSTRNG
TOKENIZE
TWORET
Routine Modifiers
ALPHANUM
DECOMP
DI
NI
Additional Routine Modifier
Compare
Routine
STREETS
Routine Modifier
Edit Multiple Routine Modifier
Score

2. Matching algorithms: families

Booleans / Classifiers	**Rules & predicates** Smals “=” Société de Mécanographie pour l’Application des Lois Sociales	**Phonetics** Dupont “=” Dubond
Similarity-based	**Word-based** Smals “=” Smlas	**Token-based** VZW Smals ASBL “=” ASBL Smals

Contents

3. Data matching with DQ tools

1. Matching: rules and patterns

2. Interpreting matching results

3. Advanced : multi-matching and transitivity

3.1. Matching: rules (deterministic approach)

- Example : setting algorithms, columns and score thresholds

Field List Editor - C:\Users\ganha\AppData\Roaming\Trillium Software\TssUI\15\cache\dmining_id_first_tests\session\E1668\project41\settings\p16_bebus1fld.stx

Description	Score A	Score B	Scor...	Scor...	Scor...	Comparison Routine	Propagati...	Field Name 1	Field Name 2	F...	Routine Modifier
business_name	99	90	80			busname		PR_BUSNAME_RECODED_01			
business_sort	95					busname		PR_BUSNAME_RECODED_01			SORT
business_squish	99					spelling		PR_BUSNAME_RECODED_01			SQUISH
business_substr	97					substrng		PR_BUSNAME_RECODED_01	PR_BUSNAME_RECODED_01		
street_name	95	89	84			streets		TS_STREET_NAME			
house_number	99	98	90			houseno		TS_HOUSE_NUMBER			
box_number_incl_blanks	95	80				aptno		PR_BOX1_NUMBER			
box_number_excl_blanks	75					partial1		PR_BOX1_NUMBER			
prevent_between_[un]...	100	0				partial1		TMP			

3.1. Matching: patterns (deterministic approach)

- The real power and flexibility of matching: patterns

- Matching pattern:
  – Combination of the rule scores we just saw
  – A pattern is either a passing, suspect or failing match

- Top to bottom
- All thresholds should be met for the pattern to pass.
- If a pattern does not pass, the next one is evaluated.
- If a Failure pattern (e.g. n°999) is hit, comparison of the two current rows stops.

Category	Pattern ID	business_name	business_sort	business_squish	business_substr	street_name	house_num...	box_number_in...	box_number_ex...	prevent_...
F	999	-	-	-	-	-	-	-	-	A
P	110	A	-	-	-	A	A	A	-	-
P	111	A	-	-	-	A	A	B	-	-
P	112	A	-	-	-	A	B	A	-	-
P	113	A	-	-	-	A	B	B	-	-
P	114	A	-	-	-	B	A	A	-	-
P	115	-	A	-	-	A	A	A	-	-
P	116	-	-	A	-	A	A	A	-	-
P	117	-	-	-	A	A	A	-	A	-
P	118	-	-	-	A	A	A	A	-	-
P	119	A	-	-	-	A	C	A	-	-
P	120	A	-	-	-	A	C	B	-	-
P	121	A	-	-	-	B	A	B	-	-
P	122	A	-	-	-	B	B	A	-	-
P	123	A	-	-	-	B	B	B	-	-
P	124	A	-	-	-	B	C	A	-	-
P	125	A	-	-	-	B	C	B	-	-
P	126	A	-	-	-	C	A	A	-	-
P	127	A	-	-	-	C	A	B	-	-
P	128	A	-	-	-	C	B	A	-	-
P	129	A	-	-	-	C	B	B	-	-
P	130	A	-	-	-	C	C	A	-	-
P	131	A	-	-	-	C	C	B	-	-
P	132	A	-	-	-	A	-	-	A	-
P	133	A	-	-	-	A	-	B	-	-
P	134	A	-	-	-	B	-	-	A	-
P	135	A	-	-	-	B	-	B	-	-
P	136	A	-	-	-	-	A	-	A	-
P	137	B	-	-	-	A	A	A	-	-
P	138	B	-	-	-	A	A	B	-	-
P	139	B	-	-	-	A	B	A	-	-
P	140	B	-	-	-	A	B	B	-	-
P	141	B	-	-	-	A	C	A	-	-
P	142	B	-	-	-	A	C	B	-	-
P	143	B	-	-	-	B	A	A	-	-
P	144	B	-	-	-	B	A	B	-	-
P	145	B	-	-	-	B	B	A	-	-
P	146	B	-	-	-	B	B	B	-	-
P	147	B	-	-	-	B	C	A	-	-
P	148	B	-	-	-	B	C	B	-	-
P	149	B	-	-	-	A	-	-	A	-
P	150	B	A	-	-	B	-	-	A	-
P	151	B	-	-	-	C	A	A	-	-
P	152	B	-	-	-	C	A	B	-	-
P	153	B	-	-	-	C	B	A	-	-
P	154	B	-	-	-	C	B	B	-	-
P	155	B	-	-	-	C	C	A	-	-
P	156	B	-	-	-	C	C	B	-	-
P	157	C	-	-	-	A	A	-	A	-
P	158	C	-	-	-	A	A	A	-	-
P	159	C	-	-	A	B	A	A	-	-
F	160	B	-	-	-	A	-	B	-	-
F	161	A	-	-	-	-	A	B	-	-
F	162	A	-	-	-	-	-	B	-	-
F	163	C	-	-	-	A	A	B	-	-
F	164	C	-	-	-	B	A	B	-	-
F	165	A	-	-	-	A	-	A	-	-

3.2. Interpreting matching results

- When two rows match, a common cluster ID (“match ID”, “group ID”, …) is generated (typically an integer ID)

- The matching pattern ID is also provided
  - Understanding why it matched
  - Fine-tuning
  - Justifying a match if needed

- Generating groups != merging data automatically (unlike OpenRefine)

3.2. Interpreting matching results

ID	NOM	ADRESSE	CODE POSTAL	LOCALITE	MATCH_ID	MATCH_PATTERN
6885	INST.ST.DOMINIQUE	RUE CAPORAL CLAES 38	1030	SCHAERBEEK	0000001053	110
6885	INST.ST.DOMINIQUE	RUE CAPORAL CLAES 38	1030	SCHAERBEEK	0000001053	110
23117	INSTITUT SAINT-DOMINIQUE	RUE CAPORAL CLAES 38	1030	BRUXELLES	0000001053	135

3.2. Interpreting matching results

ID	NOM	ADRESSE	CODE POSTAL	LOCALITE	MATCH_ID	MATCH_PATTERN
6885	INST.ST.DOMINIQUE	RUE CAPORAL CLAES 38	1030	SCHAERBEEK	0000001053	110
6885	INST.ST.DOMINIQUE	RUE CAPORAL CLAES 38	1030	SCHAERBEEK	0000001053	110
23117	INSTITUT SAINT-DOMINIQUE	RUE CAPORAL CLAES 38	1030	BRUXELLES	0000001053	135

Grade Pattern Editor - C:\Users\ganha\AppData\Roaming\Trillium Software\TssUI\15\cache\wbfin_dqm\session\E114\project5\settings\p21_

Category	Pattern ID	ID
P	110	A

Field List Editor - C:\Users\ganha\AppData\Roaming\Trillium Software\TssUI\15\cache\wbfin_dqm\session\E114\project5\settings\p21_bebu

Description	Scor...	Scor...	Scor...	Scor...	Scor...	Comparis...	Propagati...	Field Nam...	Field Nam...	Field Nam...
ID	100					partial1		NUM_BENEF		

Grade Pattern Editor - C:\Users\ganha\AppData\Roaming\Trillium Software\TssUI\15\cache\wbfin_dqm\session\E1

Category	Pattern ID	busine...	busine...	busine...	busine...	street_...	street_ext_key	house_...	city
P	135	B	-	-	-	A	-	A	-

Field List Editor - C:\Users\ganha\AppData\Roaming\Trillium Software\TssUI\15\cache\wbfin_dqm\session\E107\pr

Description	Scor...	Scor...	Scor...	Scor...	Scor...	Comparis...	Propagati...	Field Nam...	Field
business_name	100	90	80			busname		PR_BUSNAM...	
business_sort	95					busname		PR_BUSNAM...	
business_squish	100					spelling		PR_BUSNAM...	
business_substr	97					substrng		PR_BUSNAM...	PR_BU
street_name	95	87	82			streets		TS_STREET_...	
street_ext_key	100					partial1		TQ_GOUT_E...	
house_number	100	98	90			houseno		TS_HOUSE_N...	
city	99	90	85	0		spelling		TS_CITY_NAME	

3.2. Interpreting matching results : drill-down

Value	Frequency	Dist %
0000000001	3	0.016
0000000002	3	0.016
0000000003	2	0.011
0000000004	1	0.005
0000000005	1	0.005
0000000006	1	0.005
0000000007	2	0.011
0000000008	2	0.011
0000000009	1	0.005
0000000010	1	0.005

Cluster IDs

3.2. Interpreting matching results : drill-down

Value	Frequency	Dist %
0000000001	3	0.016
0000000002	3	0.016
0000000003	2	0.011
0000000004	1	0.005
0000000005	1	0.005
0000000006	1	0.005
0000000007	2	0.011
0000000008	2	0.011
0000000009	1	0.005
0000000010	1	0.005

Cluster IDs

Sort on frequency desc

Value	F...
0000008595	16
0000001858	14
0000000158	10
0000007350	10
0000012838	10
0000000053	9
0000000821	9
0000001093	9
0000006580	9
0000000148	8

Drill-down

ID	NOM	ADRESSE	CODE_POSTAL	LOCALITE	MATCH_ID	MATCH_PATTERN
0013483	INST. NOTRE DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135
0021698	INSTITUT NOTRE-DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135
0021698	INSTITUT NOTRE-DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0025572	INSTITUT NOTRE-DAME	58-68 RUE DE FIENNES	001070	BRUXELLES	0000001858	155
0039553	INSTITUT NOTRE-DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135
0039553	INSTITUT NOTRE-DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135
0039553	INSTITUT NOTRE-DAME	RUE DE FIENNES 66	001070	BRUXELLES	0000001858	135

ULB MASTIC en

3.3. Advanced : multi-matching and transitivity

General Electric International Inc.	402	29420	Dorchester Road	8550		Charleston, SC		402000073544	0000000012	217
GENERAL ELECTRIC INTERNATIONAL INC	103	60313	BLEICHSTRABE	64	66	FRANKFURT AM MAIN		103005012949	0000000012	217
General Electric	122	02256	Al. Krakowska			Warsaw		122000887589	0000000012	217
General Electric Int Inc	127	8048	Bändliweg 20			Zürich		127000159406	0000000012	217
General Electric	112	LN6 3TA	Runcorn Road	Uni4		Lincoln		112000384559	0000000012	217
General Electric International Inc.	116	17	Clonshaugh Ind Est			Dublin	0083656V	116000022676	0000000012	217
general electric	112	RG12 1PU	Diwnshre Way			Bracknell		112000384460	0000000012	217
General Electric Int Inc	112	DA1 5PZ	Littlebrook Business Park, Dartford, Kent			Kent		112000285579	0000000012	217
General Electric International Inc	116	0	Shannon Business Park			Shannon		116000149964	0000000012	217
General Electric	402	CA92705	E. Carnegie Avenue	1831		Santa Ana		402000059389	0000000012	217
General Electric International Inc.	128	20124	Via Lepetit Roberto			Milano		128000399650	0000000012	217
General Electric Int. Inc.	116	BN33FH	Clonshaugh Industrial Estate Clonshaugh	17		Dublin		116000042472	0000000012	217
General Electric International Inc	401	T6B 2L8	49th Street	9449		Edmonton		401000022251	0000000012	217
General Electric Int Inc	112	G3 8BW	2 Central Quey, Hydepark Street	89		Glasgow		112000050207	0000000012	217
General Electric	402	US12345	River Road	1		Schenectady		402000041969	0000000012	217
General Electric Int. Inc.	124	011884	Ermil Pangratti	30		Bucharest		124000012059	0000000012	217
General Electric International Inc.	111	90007	Avenue du Maréchal Juin			Belfort		111001697996	0000000012	217
GENERAL ELECTRIC INTERNATIONAL INC	611	NSW 2111	VICTORIA ROAD	450		GLADESVILLE		611000004777	0000000012	217
General Electric International Inc	109	28027	Josefa Valcárcel	26		Madrid	ESW4001041E	109000524353	0000000012	217
General Electric International Inc	109	28027	Josefa Valcarcel			Madrid		109000788926	0000000012	217
General Electric	401	V1S 2B5	Bentall Drive	2604		Kamloops		401000022152	0000000012	217
General Electric International SRL Wilmington RO	124	014459	Floreasca Road	169A		Bucharest	RO14749113	124000047790	0000000012	217
GENERAL ELECTRIC INTERNATIONAL, INC	402	CT 06828	EASTON TURNPIKE	3135		FARFIELD		402000467185	0000000012	217
General Electric	402	12345	River Road	1		Schenectady		402000024747	0000000012	217
General Electric International Wilmington S. Ro	124	041919	Berceni Road	104		Buchareast	RO14749113	124000057490	0000000012	230
General Electric International Inc.	111	90007	Avenue du Marechal Juin			Belfort		111001183007	0000000012	217
General Electric International Inc.	124	011884	Ermil Pangratti	30		Bucharest		124000012257	0000000012	217
General Electric Internation Inc.	109	28027	Jsefa Valcarce	26		Madrid		109000209993	0000000012	217
general electric	402	us 12305	1 river road			schenectady		402000328021	0000000012	217
General Electric International, Inc.	123	2774-533	Edificio D. José	1		Paco de Arcos		123000055687	0000000012	217
General Electric	402	06828	Easton Turnpike	3135		Fairfield		402000012572	0000000012	217
General Electric	112	G81 8BW	HYDEPARK STREET	89		GLASGOW		112000498286	0000000012	217
General Electric	402	NY 12345	River Road	1		Schenectady		402000312480	0000000012	217
General Electric International Inc.	112	G3 8BW	Central Quay	2		Glasgow		112000437910	0000000012	217
General Electric	111	90000	postoffice			Belfort		111000562108	0000000012	217
General Electric International Inc.	103	45141	Bamler Str.	1B		Essen	103002857965	103003801538	0000000012	217
GENERAL ELECTRIC INTERNATIONAL	112	LS1 6HP	TREVELYAN SQUARE - BOAR LANE	1		LEEDS		112000279344	0000000012	217
General Electric International	111	75009	Rue Pillet Will	2		Paris		111001281391	0000000012	217
General Electric International	111	92800	rue delarivière lefoullon - Tour défense plaza	23		Puteaux	FR21662047216	111003540602	0000000012	217
general electric	402	AA 29644	garlington road	300		greenville		402000327427	0000000012	217
GENERAL ELECTRIC	111	90000	AVENUE DES TROIS CHENES	7		BELFORT		111004899491	0000000012	217
GENERAL ELECTRIC	114	15125	Sorou	8		MAROUSI		114000018773	0000000012	217
General Electric International Inc	112	RG12 1PU	The Arena, Downshire Way	2		Bracknell		112001286263	0000000012	217
General Electric	123	2774-533	Edificio D. Jose	1		Paco de Arcos		123000053709	0000000012	217
GENERAL ELECTRIC INTERNATIONAL	111	92100	RUE DU PONT DE SEVRES	204		BOULOGNE BILLANCOURT		111006774462	0000000012	217
General Electric International Inc.	112	RG12 1PU	The Arena Downshire Way			Bracknell		112000118503	0000000012	217
GENERAL ELECTRIC INTERNATIONAL, INC.	402	GA 30339	WILDWOOD PARKWAY	4200		ATLANTA		402000337325	0000000012	217
GENERAL ELECTRIC INTERNATIONAL INC	611	NSW 2000	GEORGE STREET	255		SYDNEY		611000015368	0000000012	217
General Electric Int. Inc.	402	NJ 07047	Tonnelle ave	6001		North Bergen		402000173019	0000000012	217
General Electric International Inc.	103	60313	Bleichstraße 64-66			Frankfurt		103004811328	0000000012	217
General Electric	110	00510	Kuortaneenkatu	2		Helsinki		110000086389	0000000012	217
General Electric	112	LN6 3QP	Kingsley Trade Park			Lincoln		112000575688	0000000012	217
General Electric	112	G3 8BW	hyde park street	89		glasgow	GB531942354	112001287649	0000000012	217
GENERAL ELECTRIC INTERNATIONAL INC USA SVENSK	126	171 75	BOX 310			STOCKHOLM		126000051396	0000000012	230
General Electric Int. Inc	112	G3 8BW	2 Central Quay, 89 Hydepark Street			Glasgow		112000200160	0000000012	217
General Electric International Inc.	103	45141	Bamlerstr.	1b		Essen		103004904863	0000000012	217
General electric europe	112	WA74UH	Chandlers Court	4		Runcorn		112000056046	0000000012	217
General Electric	112	WA3 6BX	Daten Avenue			England		112000133547	0000000012	217
General Electric Int. Inc.	128	20126	via chiese	72		milan		128000141908	0000000012	217
General Electric International Inc.	402	TX 77346	12226 Salt River Valley Lane			Humble TX		402000245966	0000000012	217
General Electric International Inc.	402	AA 60527	7521 Brush Hill Road			Burr Ridge Illinois		402000330001	0000000012	217

3.3. Advanced : multi-matching and transitivity : results across 3 DBs

- Zoom from previous slide

General Electric International Inc.	116	17	Clonshaugh Ind Est			Dublin	0083656V
General Electric Int. Inc.	116	BN33FH	Clonshaugh Industrial Estate Clonshaugh	17		Dublin	
General Electric	112	G81 8BW	HYDEPARK STREET	89		GLASGOW	
General Electric Int. Inc	112	G3 8BW	2 Central Quay, 89 Hydepark Street			Glasgow	
GENERAL ELECTRIC INTERNATIONAL INC	103	60313	BLEICHSTRABE	64	66	FRANKFURT AM MAIN	
General Electric International Inc.	103	60313	Bleichstraße 64-66			Frankfurt	

Contents

4. Performance and blocking / windowing

- Matching = comparisons. Naive approach:

$$N = n^2$$

- (A bit) Less naive approach: not comparing R with itself

$$N = n^2 - n$$

  - n: number of rows to compare
  - N: total number of comparisons

4. Performance and blocking / windowing

- Optimal approach (comparisons are not directional):

$$N = \frac{n^2 - n}{2}$$

  - n: number of rows to compare
  - N: total number of comparisons

- Thus time complexity remains ~ $O(n^2)$

4. Performance and blocking / windowing

- N comparisons in practice
  - $n = 3\,rows \rightarrow N = 3$
  - $n = 6 \rightarrow N = 15$
  - $n = 10\,000\,000 \rightarrow N = 49\,999\,995\,000\,000$

- For each comparison (pair of rows)
  - $*$ p: from one up to dozens of **patterns** to test
    - $*$ a: multiple **attributes** to process per pattern per row
      - $*$ t: from one up to dozens of **transformations** per attribute (comparison algorithms)

80999991900000000

80 999 991 900 000 000

4. Performance and blocking / windowing

- In total, $Np * (2at + 1)$ logical operations
  - For 10 million rows:

$$Np * (2at + 1) = 80\,999\,991\,900\,000\,000$$

- Assuming a common situation where $p = 20$, $a = 4$, and $t = 10$.

4. Performance and blocking / windowing: principle

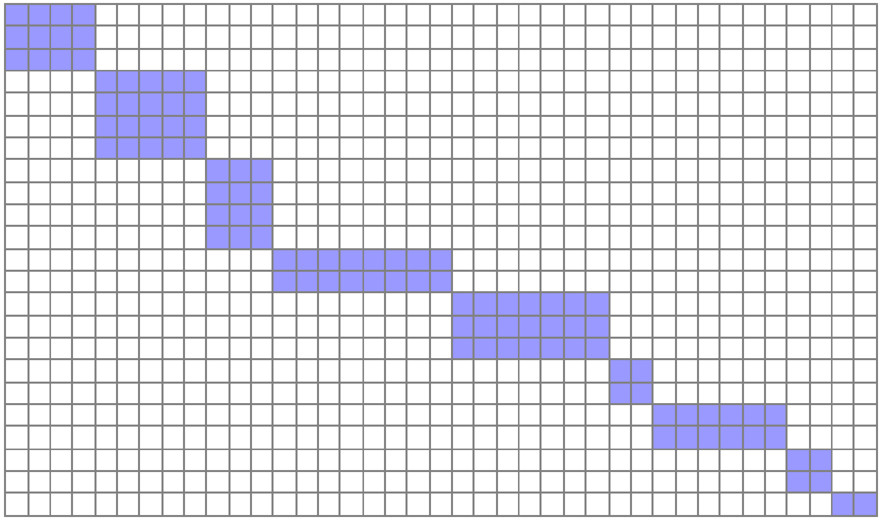
- Derive « **keys** » to split data rows in **subgroups (windows)**
  - /!\ quality of the source attributes that are used → business !

- E.g. key based on 1 rule (many other possible choices):
  - 4 char Soundex*(T_NAME_SRCE)
  - Needs to be tuned iteratively

T Name Srce	T Street Srce	C Zipcode Srce	T City Srce	Window Key 01
STAD ROESELARE	BOTERMARKT	8800	ROESELARE	**S336**

4. Performance and blocking / windowing: principle

- Comparisons for matching happen **only within each window** (e.g., here in 2 dimensions)



- Size of each window = determinant for feasibility
  - Windows around 500 to 1000 records are a sweet spot
  - Time performance vs. completeness (recall) of matching

4. Performance and blocking / windowing: Multimatching

- Using multiple window keys per dataset to improve matching results
  - Several data sources
  - Several window keys per source
    - Eg, for enterprises : postal code, NACE code (activity category), etc.
  - Several matching processes per source
  - Several matching processes between sources over time

T Name Srce	T Street Srce	C Zipcode Srce	T City Srce	Window Key 01	Window Key 02
STAD ROESELARE	BOTERMARKT	8800	ROESELARE	**S336**	**88B365**
ROESELARE STAD	BOTER MARKT	8800	ROESELARE	**R246**	**88B365**

4. Performance and blocking / windowing: performance gains

- Finding appropriate window keys requires analysis and iterations. Worth it.

- After optimizing window keys for the flow we just saw:
  - 59 processes, >6 million rows * 49 attributes avg, 10GB initial srce, 4 matching processes

Total elapsed time	178:04:45

→

Total elapsed time	24:02:58

REF_MATCH_NAME_NOK_ADDRESS_NOK	rellink
Details	
Output	6.02 million rows (total data
Elapsed Time	91:44:57

→

REF_MATCH_NAME_NOK_ADDRESS_NOK	rellink
Details	
Output	6.02 million rows (total)
Elapsed Time	03:25:54

Contents

5. Golden record / survivorship

5. Golden record / survivorship

- If deduplication is needed, it is possible to build a “golden record”
- The “golden record” is the result of the best parts of each record in a matching group
- Chosing parts of different records is called commonization
- /!\ if deduplication → keep history of previous records!

5. Golden record / survivorship

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

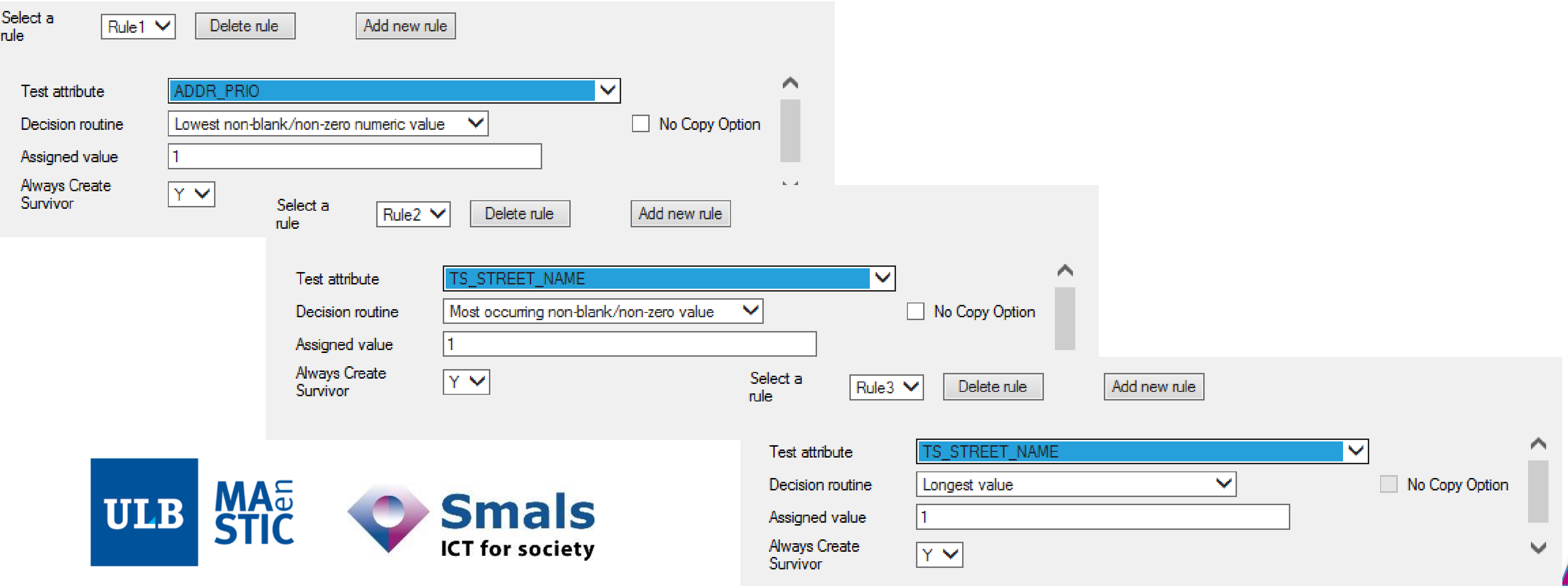
12/2/2005	Arthur	Barrios	908-845-1234	abarrios@accen.com

ULB MASTIC Smals ICT for society

Source: Y. Bontemps, D. Van Dromme

5. Golden record / survivorship Example of real rules

- Address : the most valid
  - Totally valid
    - If not : problem in house number
      - If not : problem in street name
        - » If not : problem in city name
  - If tied : the most frequent address
  - If still tied : the longest



5. Golden record / survivorship

- Documenting the decisions is key and sometimes even required by law
  - Eg Registre National (NISS and BISS number) ; source : Isabelle Boydens

"BAUDOUIN, Roi des Belges,
A tous présents et à venir, Salut.
[...]
Vu l'urgence;
Art. 5 Si le jour ou le mois de naissance d'une personne ne sont pas connus, la date de naissance est composée comme suit : [...]
Si l'année de naissance d'une personne n'est pas connue, [...]
Art. 6 Un numéro d'identification qui a déjà été utilisé ne peut être attribué à nouveau ni avant qu'un délai de cent ans ne se soit écoulé depuis la date de naissance du titulaire précédent, ni avant que celui-ci soit décédé depuis trente ans au moins.
[...]
Art 8. Si deux ou plusieurs numéros d'identification sont attribués à une même personne, un seul numéro d'identification est retenu. Les autres numéros sont détruits. Pour déterminer le numéro retenu, il est donné priorité, en ordre décroissant, au :
- numéro d'identification attribué conformément à l'arrêté royal du 3 avril 1984 relatif à la composition du numéro d'identification des personnes inscrites au Registre national des personnes physiques.
- numéro d'identification attribué en exécution du présent arrêté, dont on ne peut déduire la date de naissance, ou une partie de celle-ci, ainsi que le sexe;
- numéro d'identification attribué en exécution du présent arrêté, dont on peut uniquement déduire la date de naissance ou une partie de celle-ci;
- numéro d'identification attribué en exécution du présent arrêté, dont on peut uniquement déduire le sexe;
- numéro d'identification attribué en exécution du présent arrêté, ayant le numéro d'ordre le plus élevé.

Art. 9. Un numéro d'ordre attribué conformément au présent arrêté n'est pas modifié lorsque, après attribution du numéro, les données y reprises relatives à la date de naissance ou au sexe de la personne s'avèrent inexactes [...]".

Arrêté royal du 8/02/91 relatif à la composition et aux modalités d'attribution du numéro d'identification des personnes physiques qui ne sont pas inscrites au Registre National des personnes physiques. *Moniteur belge*, 19 février 1991.

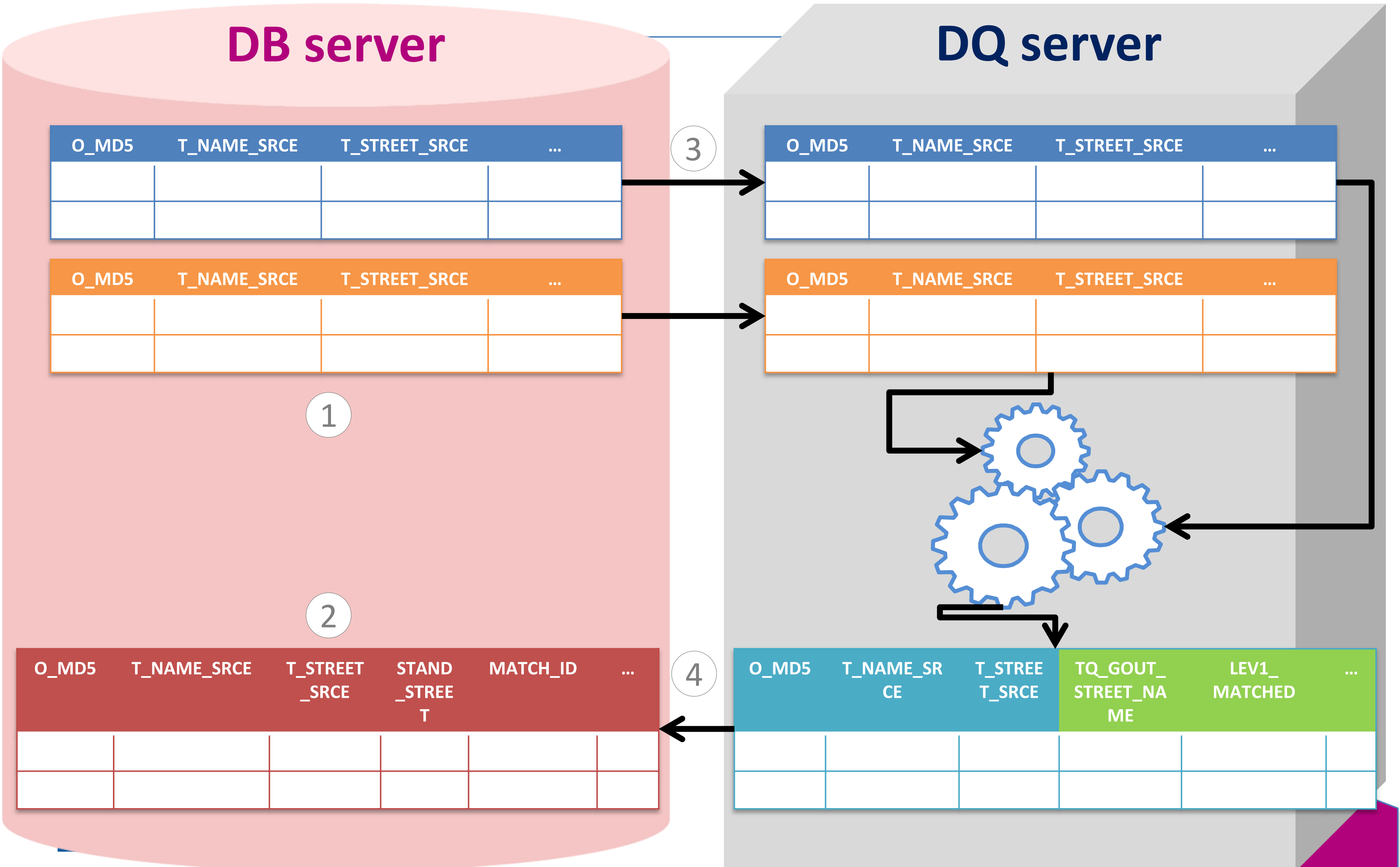
Contents

6. Running a project

- Spreadsheet-like approach:
  - Apply modifications in-place
  - Export modified dataset
  - Export modification script for later re-use

- Data flow approach:
  - Build the project through a GUI client
  - Run the project / a sample through the client
  - Export to batch (typically : big piece of Java / Bash / … code)
  - Possibly schedule runs / wait for third party server orders

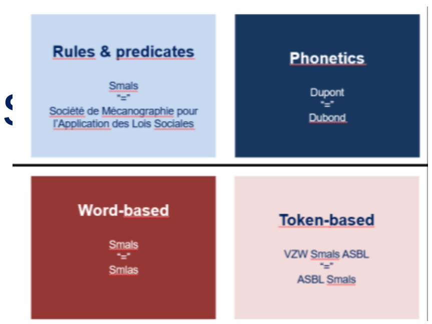
6. Running a project: DB read/write in batch



Contents

Conclusion

- To summarize:
  - Technical approach, very evolutive
  - Profiling: data & metadata audit
  - Standardization
    - Parsing
    - Validation & enrichment for some fields
  - Matching and optimizing performance
    - 4 algo (eg. **Soundex**, **Levenshtein**) families
    - Blocking
    - Golden record (w/ business)



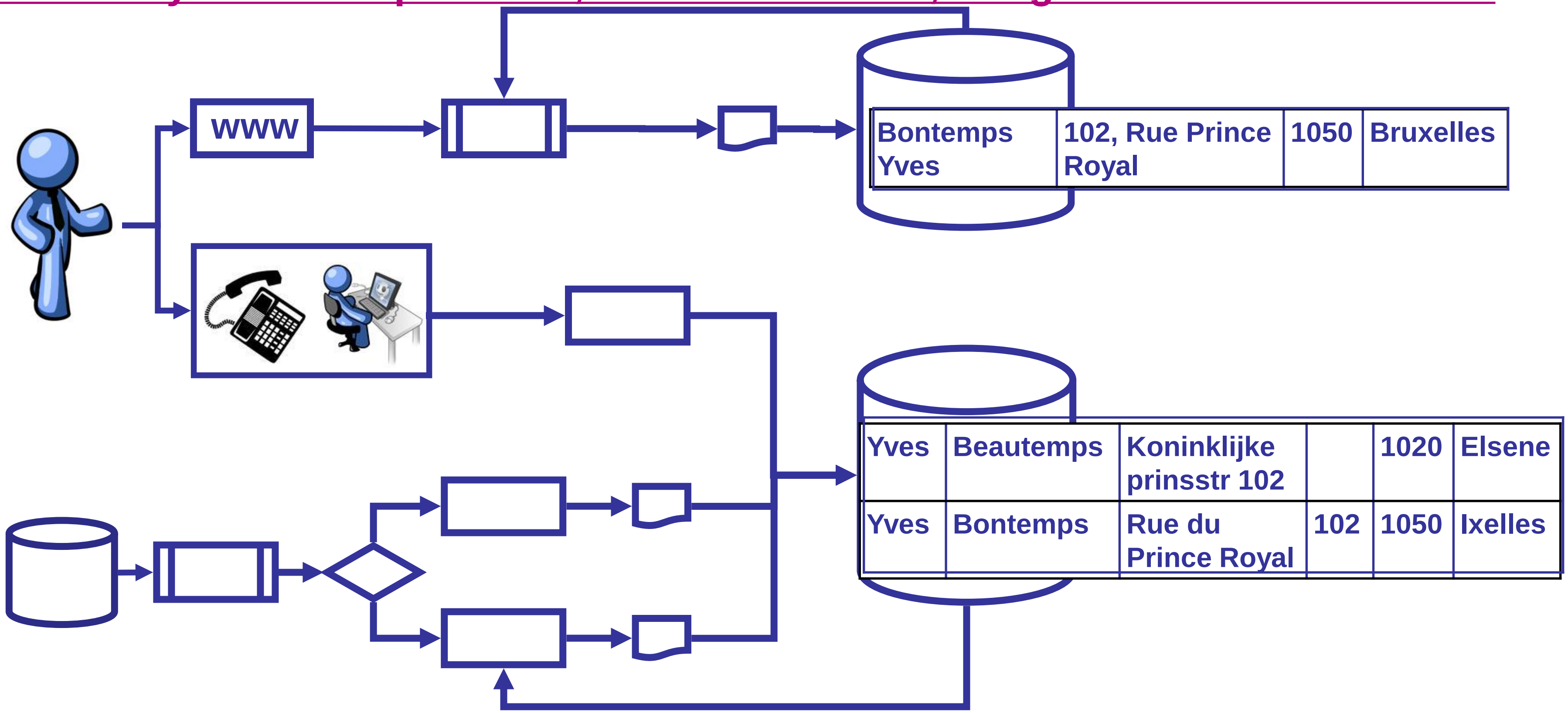
ULB MA en STIC Smals ICT for society

Conclusion

- General takeaways
  - Never-ending iteration
    - Business owners
    - Methodological approach: going to the source of DQ problems
  - When DQ has strategic impact
    - Changing data usage (migration, integration ; evolving anomalies)
    - Business inefficiency
    - Costs
  - Decades of optimizations + performance =
    - Focusing on logic instead of code
    - Dealing with huge datasets in reasonable timespans
    - Easy collab

Conclusion : two complementary approaches – continuity and recursivity

First: identify business priorities, «fitness for use», budget and «cost-benefits»



Preventive approaches **Curative approaches**

ULB MAen STIC Smals ICT for society

Conclusion

- Future: “machine learning” or other technical approaches
  - No big breakthrough yet
    - Some tools (e.g. Talend) offer basic ML functionalities
  - Caution around
    - Operational results involving real and validated business case studies
    - The “explainability” of results

Documentation

Boydens I., *Informatique, normes et temps*, Bruylant, 1999.

Smals Research (Isabelle Boydens, Yves Bontemps, Dries Van Dromme) about data quality & DQ tools

- Gestion intégrée des anomalies
  - https://www.smalsresearch.be/?wpfb_dl=62
- Data quality tools :
  - https://www.smalsresearch.be/?wpfb_dl=85

Olson J., *Data Quality: the Accuracy Dimension*. Elsevier: The Morgan-Kaufmann Series in Database Management, 2002