

# **Data Linkage in IDM Systems**

**Ms.C WIng, Ms.C. CE Krasimir Stoyanov Zhelev**  
**Projects & Processes – IDMone**  
**20. February 2008**

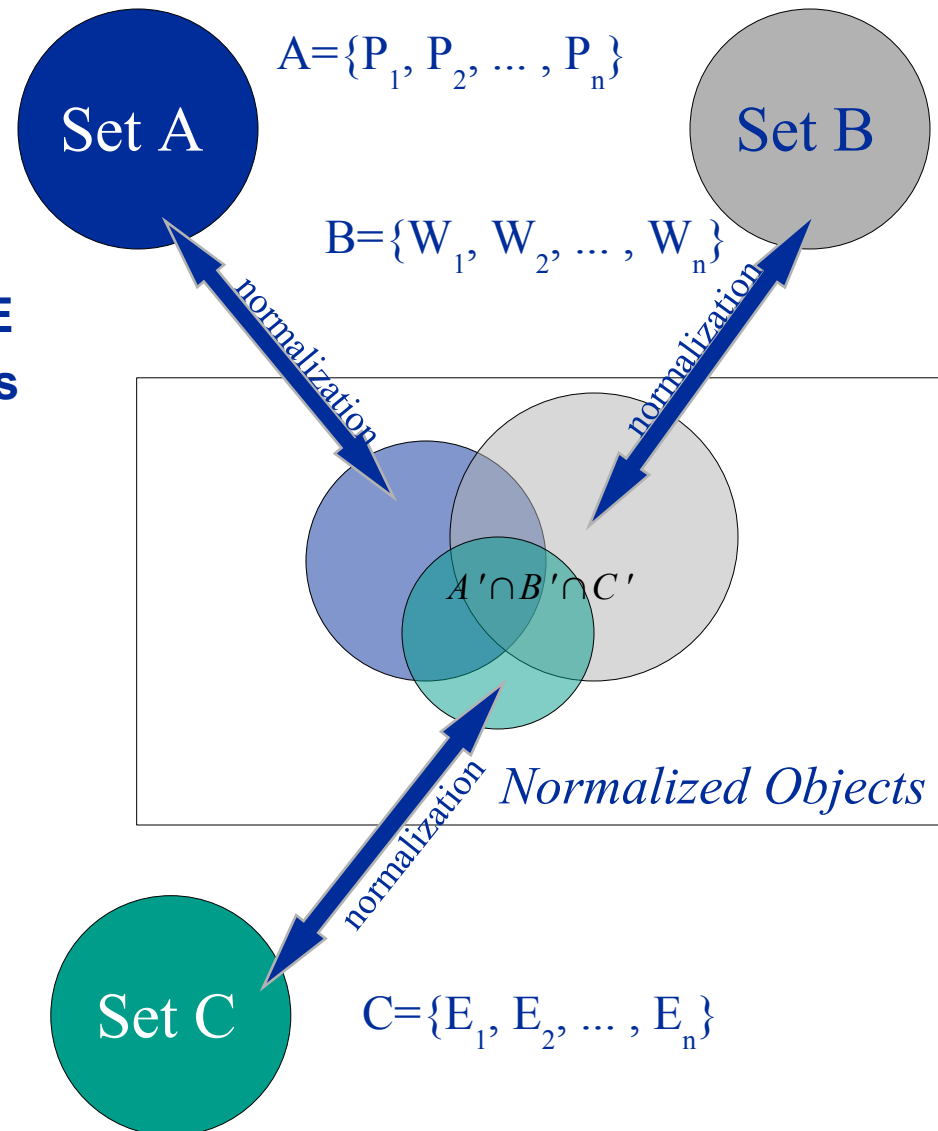


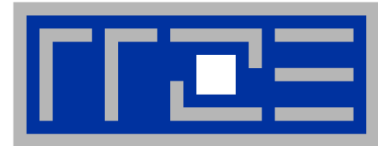
- **Data Linkage Systems - Overview**
- **Normalization**
- **Data Sets**
- **Data Linkage Specifics**
- **Name matching**
- **Similarity Functions**
- **Tuning and Refinements**
- **Web-based Reporting Tool**
- **Results**
- **Conclusion**



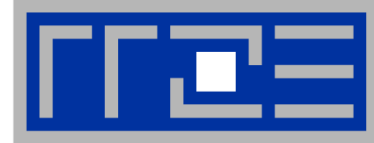
- **Goal: Linking and aggregating data form various sources that refers to the same entity**
- **Reasons:**
  - de-duplication of databases
  - improve data quality
  - ensure data integrity
  - extend existent data
  - provide basis for statistical evaluations
  - support data mining
- **Problems:**
  - no unique identifiers – attributes matching have to be used
  - classification of object matches, non-matches and possible matches has to be performed
  - large amounts of data should be processed
  - different types and formats of attributes have to be compared
  - comparison can be computationally expensive
  - automation is not feasible

- Reasoning
  - Different types of objects:
    - Persons – Object P
    - Affiliations – Object W
    - Entitlements – Object E
  - Different types of attributes
  - Different value formats
    - Dates
    - Names
  - Mappings cardinality
    - one-to-one
    - one-to-many
    - many-to-one
  - Data consistency
    - same semantics
    - same format





- **Ontology**
  - **Completeness Rule**
    - as many attributes should be mapped as possible
    - allows cross system mappings
  - **Clarity Rule**
    - **Semantic definition of a Normalized Object(NO)**
    - **Representation:**  $NO = \{A_1, A_2, \dots, A_n\}$ 
      - usually by extending an existent type
      - proper attribute types should be selected
    - **Attributes set definition**
      - type – string, date, number
      - value – format and normed form (umlaut conversion)
    - **Constraint definitions**
      - imposed on the value of an attribute
      - related to the semantic meaning of the attribute
      - garbage data collection – dates, name



## Ontology Overlapping

NO	ID	Source	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>n</sub>
A'	121525	sos	Yes	Yes	...	No
B'	2118945	diapers	Yes	Yes	...	Yes
C'	21334	legacy	No	No	...	Yes

## Weighted Overlapping

NO	ID	Source	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>n</sub>
A'	121525	sos	0.9	0.75	...	0
B'	2118945	diapers	0.85	0.87	...	0.96
C'	21334	legacy	0	0	...	0.96

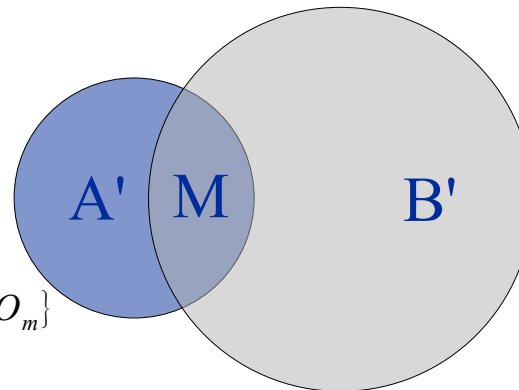
## Case review

### Typical case

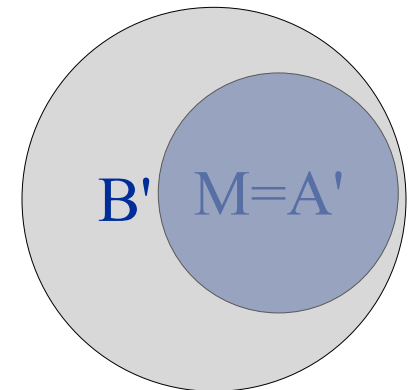
$$M = A' \cap B' = \{NO_1, NO_2, \dots, NO_m\}$$

### Containment case

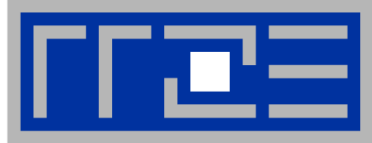
$$M = A'; M = A' \cap B' = \{NO_1, NO_2, \dots, NO_m\}$$



Typical



Containment



- **Error Sets:**
  - false negatives –  $M_{A'B'}$
  - false positives:
    - from set A –  $M_{A'}$
    - from set B –  $M_{B'}$

- **Total Number of Sets:**

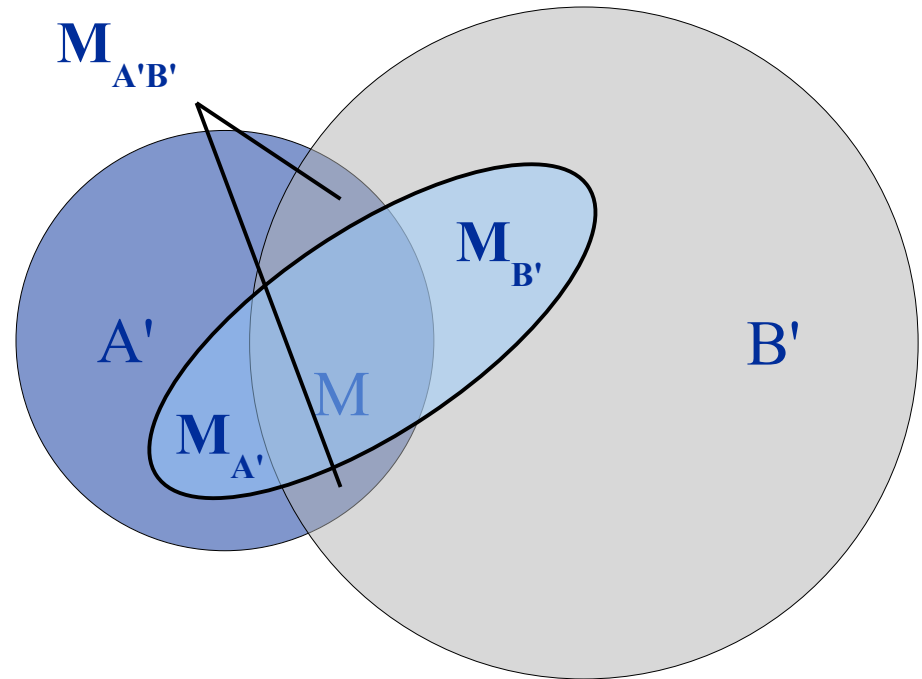
$$N = n^n + 1 = 2^2 + 1 = 5$$

- **Number of Subsets:**

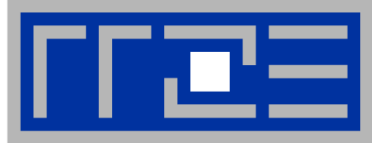
$$SN = n^{n-1} + 1 = 2^{2-1} + 1 = 3$$

- **Number of Error Sets:**

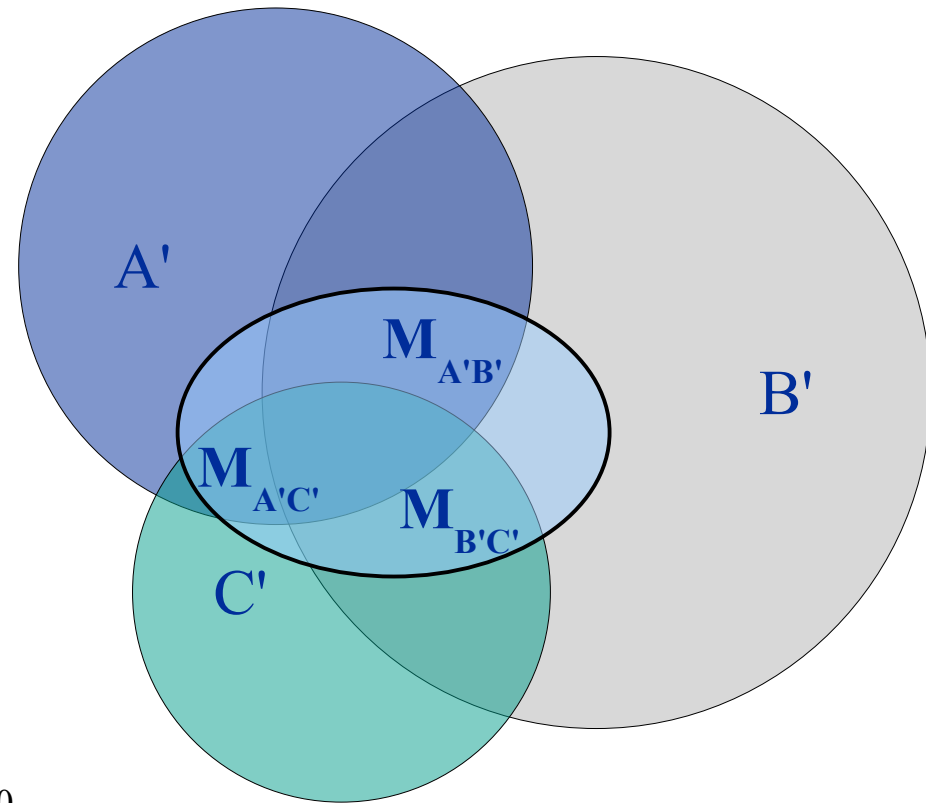
$$ES = n^{n-1} + 1 = 2^{2-1} + 1 = 3$$



Errors in two sets  
matching



- **Error Sets:**
  - false positives:
    - from sets A&B:  $M_{A'B'}$
    - from sets A&C:  $M_{A'C'}$
    - from sets B&C:  $M_{B'C'}$
  - false positives:
    - from set A:  $M_{A'}$
    - from set B:  $M_{B'}$
    - from set C:  $M_{C'}$



Errors in three sets matching

- **Sets Total:**  $N = n^n + 1 = 3^3 + 1 = 10$
- **Subsets:**  $SN = n^{n-1} + 1 = 3^{3-1} + 1 = 7$
- **Error Sets:**  $ES = n^{n-1} + 1 = 3^{3-1} + 1 = 7$



- **Minimization problem defined on the error sets**
- **Data does not match perfectly**
- **Research shows: 80% of attribute errors are single errors**
- **Most common error types:**
  - 1) **A letter was substituted for another letter**
  - 2) **A letter is deleted**
  - 3) **An extra letter is inserted**
  - 4) **Two adjacent letters are transposed**
- **Errors according to data source**
  - **OCR – similar looking characters or sequences**
  - **keyboard – neighboring keys**
  - **telephone – assuming spelling**
  - **system limitations – max. length of input field**
  - **human factor – different reporting of data**
- **Different sources match worse**



- **Identifying and linking personal data is part of IDM**
- **Most important person related linkage attributes:**
  - **Name – first name, surname**
  - **Date of birth**
  - **Place of birth**
  - **Address**
- **Generally there is no legislation on naming conventions**
- **Names have no correct spelling but rather a set of legitimate name variations**
- **Common problems:**
  - **Different spelling – Meier, Meyer, Maier**
  - **Different structure – middle name (Stoyanov)**
  - **Nicknames, short names – (Karl - Charlie, Wilhelm - Willi)**
  - **Names change – getting married, real name change**
  - **Compound names - (Hans-Peter)**
  - **Different transliterations – (Krassimir, Krasimir)**



- **Phonetic Encoding**
  - Soundex – keeps first letter encodes the others
  - Phonet – improved German version of Soundex
  - Phonix – different rules for start, middle, end of word
  - ...
- **Pattern Matching**
  - Levenshtein – counts insertions, deletions and substitutions
  - Damerau-Levenshtein Distance – includes transpositions
  - Smith-Waterman – developed for DNA sequences, match scores
  - Jaro – also estimates transpositions
  - Jaro-Winkler – empirically improved Jaro for start of word
  - ...
- **Combined**
  - Editex
  - Syllable Alignment Distance
  - ...



- **Similarities as stored procedures**
  - prevents loading of large amounts of data in memory
  - makes easier and faster constructing filter queries
  - all functions give estimates from 0.0 to 1.0

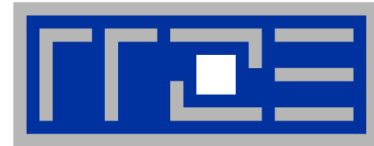
```
SIMILARITY_PLACEHOLDER( a.norm_surname,  
b.norm_surname ) > VALUE_PLACEHOLDER
```

- **Frequency Distribution of names**
  - Calculated from the data sets

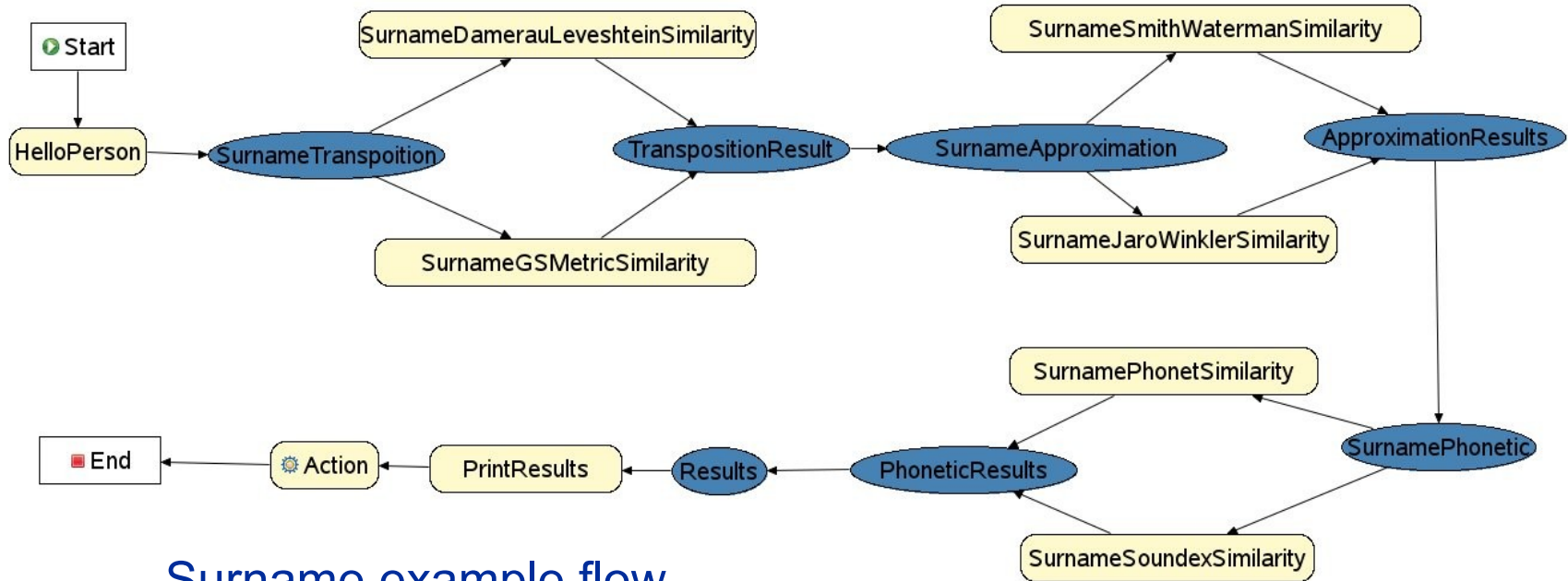
```
RRZELinkage.PMMatchGivenname(varchar,varchar)
```

```
RRZELinkage.PMMatchCitizenship(varchar,varchar)
```

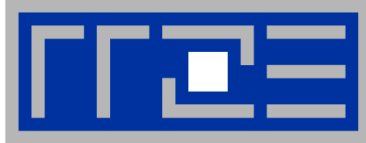
- **Calculated from complete database of names**



- Integrating a Business Rules Engine allows:
  - implementing a more complicated matching logic
  - investigating which combinations of similarities is optimal
  - customization of the obtain results
  - customization of initial filter queries
  - appropriately handling running system (not only initial load)



Surname example flow



- main
- Administration section
- install
- import
- normalize
- remove
- Public section
- display
- report
- ruleengine



## Display

### Results

### Reports

#### Matching

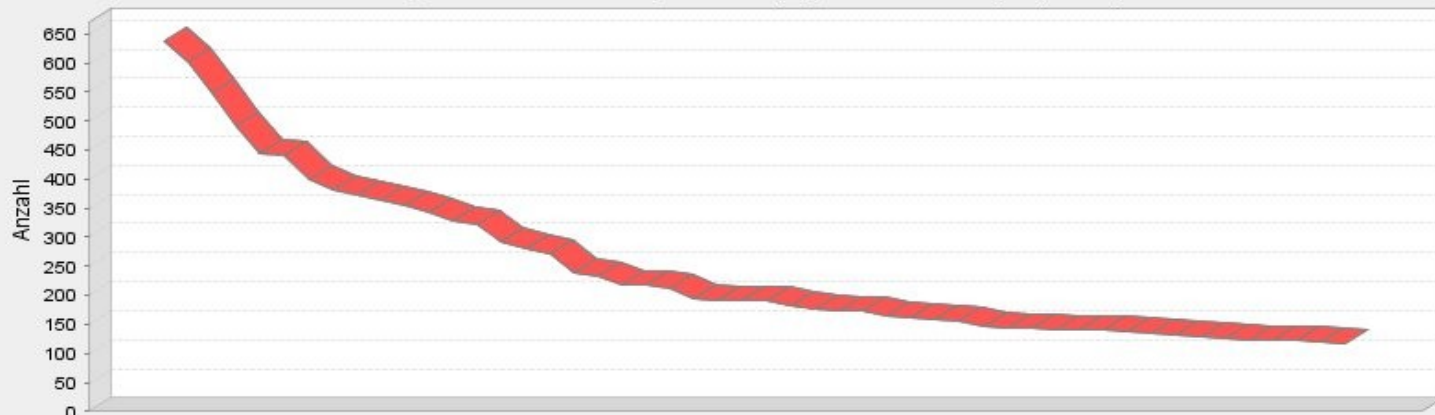
- [Exact Matching after Normalization](#)

#### Cases for clarification

- [Normalization Errors](#)
- [Double Entries](#)
- [Typos in Name](#)
- [Typos in Date of Birth](#)
- [Typos in Gender](#)
- [Single Source Diapers](#)
- [Single Source SOS](#)
- [All Reports](#)

### Frequency Distributions

Häufigkeitsverteilung - norm\_givenname (Top-50)





main

Administration section

install

import

normalize

remove

Public section

display

report

ruleengine

### Reports

Format  
 HTML  Excel


Types

- diapers only
- matches
- typos in name
- double entries
- typos in gender
- typos in date of birth
- sos only

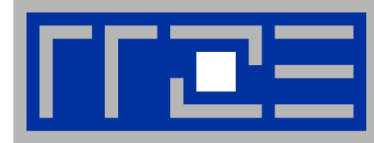
Threshold

Similarities

**Submit Query**

A logo consisting of three interlocking gears. The top gear has the letter 'I', the middle gear has 'D', and the bottom gear has 'M'. To the right of the gears, the word 'Done' is written in a blue, stylized font.

[Nach oben](#)




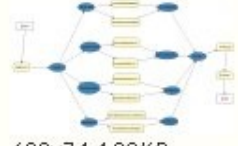
## Rule Flow Management

Browse

[RuleFile](#) [RuleFlow](#) [SubFlow](#)

[home](#) [browse](#) [add](#)

Pages: 1 ( << < > >> | 1 )

moment	name	image	column_label-actions
19.02.2008 12:07:31	RefinedRuleFlow	 120x80 7.54KB	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
13.02.2008 15:05:06	MyFirstRuleFlow	 120x74 4.09KB	<input type="button" value="Edit"/> <input type="button" value="Delete"/>



- **Data linkage is a complex and error prone process**
- **Gained experience so far:**
  - **It is important to know the specifics of the involved systems.**
  - **First fast approximation functions should be used to filter out possible negative positives.**
  - **Phonetic comparison should always be combined with an approximation function unless specifically searching for phonetic errors.**
  - **Smith-Waterman approximation provides better results when compared to other similarity methods.**
  - **Significant effort should be allocated to tuning up thresholds.**
  - **rule based engine can be used to improve results.**
- **A framework is developed to allow the generation of various reports and testing of different scenarios**
- **Future work:**
  - **derive suitable rule-flows for different matching purposes**
  - **better utilization of frequency distributions**



**Thank You for the attention!**



## Krasimir Zhelev

Entwickler Identity Management

RRZE Martensstrasse 1

D-91058 Erlangen

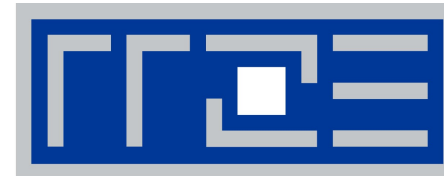
Tel.: +49 9131 85-28145

Fax: +49 9131 302941

[krasimir.zhelev@rrze.uni-erlangen.de](mailto:krasimir.zhelev@rrze.uni-erlangen.de)

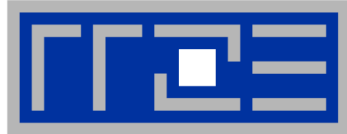
<http://www.rrze.uni-erlangen.de>

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)



Regionales  
Rechenzentrum  
Erlangen

Der IT-Dienstleister der FAU



## **Data Linkage in IDM Systems**

**Ms.C WIng, Ms.C. CE Krasimir Stoyanov Zhelev**  
**Projects & Processes – IDMone**  
**20. February 2008**



- **Data Linkage Systems - Overview**
- **Normalization**
- **Data Sets**
- **Data Linkage Specifics**
- **Name matching**
- **Similarity Functions**
- **Tuning and Refinements**
- **Web-based Reporting Tool**
- **Results**
- **Conclusion**

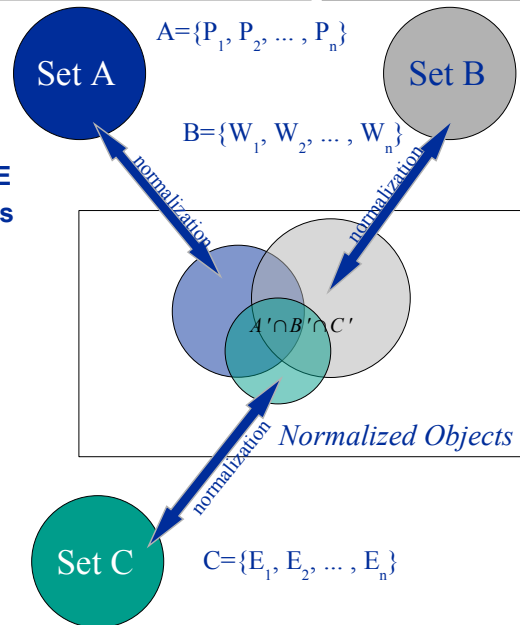


- **Goal: Linking and aggregating data form various sources that refers to the same entity**
- **Reasons:**
  - de-duplication of databases
  - improve data quality
  - ensure data integrity
  - extend existent data
  - provide basis for statistical evaluations
  - support data mining
- **Problems:**
  - no unique identifiers – attributes matching have to be used
  - classification of object matches, non-matches and possible matches has to be performed
  - large amounts of data should be processed
  - different types and formats of attributes have to be compared
  - comparison can be computationally expensive
  - automation is not feasible

# Normalization



- Reasoning
  - Different types of objects:
    - Persons – Object P
    - Affiliations – Object W
    - Entitlements – Object E
  - Different types of attributes
  - Different value formats
    - Dates
    - Names
  - Mappings cardinality
    - one-to-one
    - one-to-many
    - many-to-one
  - Data consistency
    - same semantics
    - same format





- **Ontology**
  - **Completeness Rule**
    - as many attributes should be mapped as possible
    - allows cross system mappings
  - **Clarity Rule**
    - **Semantic definition of a Normalized Object(NO)**
    - **Representation:**  $NO = \{A_1, A_2, \dots, A_n\}$ 
      - usually by extending an existent type
      - proper attribute types should be selected
    - **Attributes set definition**
      - type – string, date, number
      - value – format and normed form (umlaut conversion)
    - **Constraint definitions**
      - imposed on the value of an attribute
      - related to the semantic meaning of the attribute
      - garbage data collection – dates, name



## Ontology Overlapping

NO	ID	Source	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>n</sub>
A'	121525	sos	Yes	Yes	...	No
B'	2118945	diapers	Yes	Yes	...	Yes
C'	21334	legacy	No	No	...	Yes

## Weighted Overlapping

NO	ID	Source	A <sub>1</sub>	A <sub>2</sub>	...	A <sub>n</sub>
A'	121525	sos	0.9	0.75	...	0
B'	2118945	diapers	0.85	0.87	...	0.96
C'	21334	legacy	0	0	...	0.96

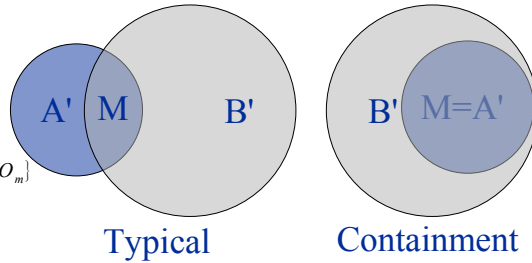
## Case review

### Typical case

$$M = A' \cap B' = \{NO_1, NO_2, \dots, NO_m\}$$

### Containment case

$$M = A'; M = A' \cap B' = \{NO_1, NO_2, \dots, NO_m\}$$



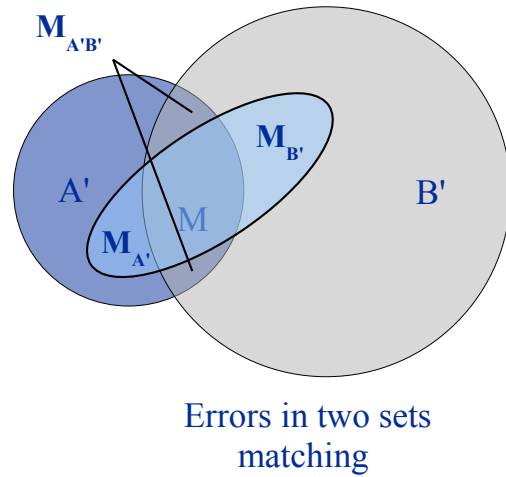


- **Error Sets:**
  - false negatives –  $M_{A'B'}$
  - false positives:
    - from set A –  $M_{A'}$
    - from set B –  $M_{B'}$
- **Total Number of Sets:**

$$N = n^n + 1 = 2^2 + 1 = 5$$
- **Number of Subsets:**

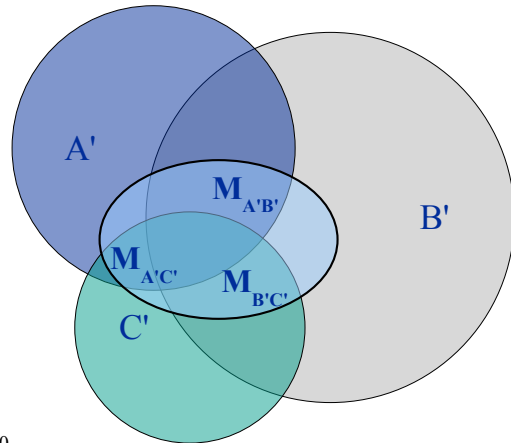
$$SN = n^{n-1} + 1 = 2^{2-1} + 1 = 3$$
- **Number of Error Sets:**

$$ES = n^{n-1} + 1 = 2^{2-1} + 1 = 3$$





- **Error Sets:**
  - false positives:
    - from sets A&B:  $M_{A'B'}$
    - from sets A&C:  $M_{A'C'}$
    - from sets B&C:  $M_{B'C'}$
  - false positives:
    - from set A:  $M_{A'}$
    - from set B:  $M_{B'}$
    - from set C:  $M_{C'}$
- **Sets Total:**  $N = n^n + 1 = 3^3 + 1 = 10$
- **Subsets:**  $SN = n^{n-1} + 1 = 3^{3-1} + 1 = 7$
- **Error Sets:**  $ES = n^{n-1} + 1 = 3^{3-1} + 1 = 7$



Errors in three sets  
matching



- **Minimization problem defined on the error sets**
- **Data does not match perfectly**
- **Research shows: 80% of attribute errors are single errors**
- **Most common error types:**
  - 1) **A letter was substituted for another letter**
  - 2) **A letter is deleted**
  - 3) **An extra letter is inserted**
  - 4) **Two adjacent letters are transposed**
- **Errors according to data source**
  - **OCR – similar looking characters or sequences**
  - **keyboard – neighboring keys**
  - **telephone – assuming spelling**
  - **system limitations – max. length of input field**
  - **human factor – different reporting of data**
- **Different sources match worse**

## Name matching



- **Identifying and linking personal data is part of IDM**
- **Most important person related linkage attributes:**
  - **Name – first name, surname**
  - **Date of birth**
  - **Place of birth**
  - **Address**
- **Generally there is no legislation on naming conventions**
- **Names have no correct spelling but rather a set of legitimate name variations**
- **Common problems:**
  - **Different spelling – Meier, Meyer, Maier**
  - **Different structure – middle name (Stoyanov)**
  - **Nicknames, short names – (Karl - Charlie, Wilhelm - Willi)**
  - **Names change – getting married, real name change**
  - **Compound names - (Hans-Peter)**
  - **Different transliterations – (Krassimir, Krasimir)**

## Similarity Functions



- **Phonetic Encoding**
  - Soundex – keeps first letter encodes the others
  - Phonet – improved German version of Soundex
  - Phonix – different rules for start, middle, end of word
  - ...
- **Pattern Matching**
  - Levenshtein – counts insertions, deletions and substitutions
  - Damerau-Levenshtein Distance – includes transpositions
  - Smith-Waterman – developed for DNA sequences, match scores
  - Jaro – also estimates transpositions
  - Jaro-Winkler – empirically improved Jaro for start of word
  - ...
- **Combined**
  - Editex
  - Syllable Alignment Distance
  - ...

20.02.08

krasimir.zhelev@rrze.uni-erlangen.de

Data Linkage in IDM Systems

11

## Tuning and Refinement



- **Similarities as stored procedures**
  - prevents loading of large amounts of data in memory
  - makes easier and faster constructing filter queries
  - all functions give estimates from 0.0 to 1.0

```
SIMILARITY_PLACEHOLDER( a.norm_surname,  
b.norm_surname ) > VALUE_PLACEHOLDER
```

- **Frequency Distribution of names**
  - Calculated from the data sets

```
RRZELinkage.PMMatchGivenname(varchar,varchar)
```

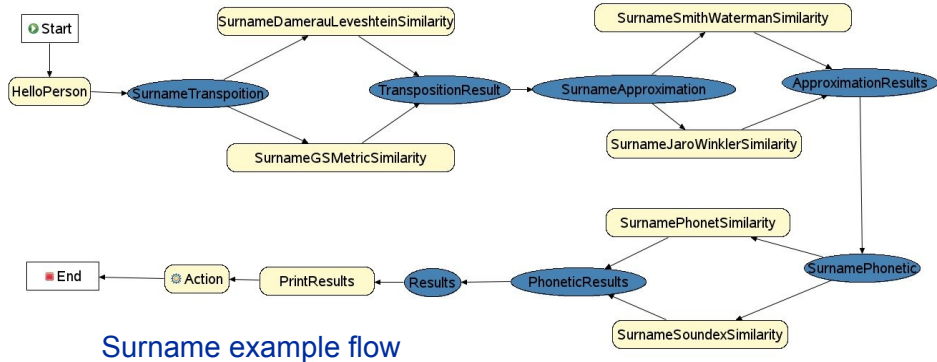
```
RRZELinkage.PMMatchCitizenship(varchar,varchar)
```

- **Calculated from complete database of names**

## Tuning and Refinement II



- Integrating a Business Rules Engine allows:
  - implementing a more complicated matching logic
  - investigating which combinations of similarities is optimal
  - customization of the obtain results
  - customization of initial filter queries
  - appropriately handling running system (not only initial load)



Surname example flow



- main
- Administration section
  - install
  - import
  - normalize
  - remove
- Public section
  - display
  - report
  - ruleengine



## Display

### Results

### Reports

### Matching

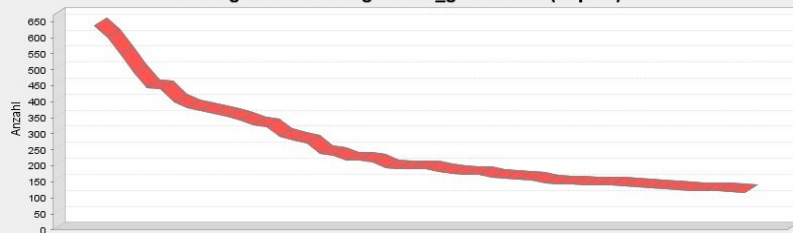
- [Exact Matching after Normalization](#)

### Cases for clarification

- [Normalization Errors](#)
- [Double Entries](#)
- [Typos in Name](#)
- [Typos in Date of Birth](#)
- [Typos in Gender](#)
- [Single Source Diapers](#)
- [Single Source SOS](#)
- [All Reports](#)

### Frequency Distributions

Häufigkeitsverteilung - norm\_givenname (Top-50)






main

Administration section

- install
- import
- normalize
- remove

Public section

- display
- report
- ruleengine



### Reports

Format  
 HTML  Excel

Types

- diapers only
- matches
- typos in name
- double entries
- typos in gender
- typos in date of birth
- sos only

Threshold

Similarities

Nach oben





### Rule Flow Management

Browse

RuleFile RuleFlow SubFlow

home browse add

Pages: 1 ( << >> | 1 )

moment	name	image	column_label-actions
19.02.2008 12:07:31	RefinedRuleFlow	 120x80 7.54KB	<input type="button" value="Edit"/> <input type="button" value="Delete"/>
13.02.2008 15:05:06	MyFirstRuleFlow	 120x74 4.09KB	<input type="button" value="Edit"/> <input type="button" value="Delete"/>

## Conclusion



- **Data linkage is a complex and error prone process**
- **Gained experience so far:**
  - **It is important to know the specifics of the involved systems.**
  - **First fast approximation functions should be used to filter out possible negative positives.**
  - **Phonetic comparison should always be combined with an approximation function unless specifically searching for phonetic errors.**
  - **Smith-Waterman approximation provides better results when compared to other similarity methods.**
  - **Significant effort should be allocated to tuning up thresholds.**
  - **rule based engine can be used to improve results.**
- **A framework is developed to allow the generation of various reports and testing of different scenarios**
- **Future work:**
  - **derive suitable rule-flows for different matching purposes**
  - **better utilization of frequency distributions**

20.02.08

krasimir.zhelev@rrze.uni-erlangen.de

Data Linkage in IDM Systems

17



**Thank You for the attention!**



**Krasimir Zhelev**

Entwickler Identity Management

RRZE Martensstrasse 1

D-91058 Erlangen

Tel.: +49 9131 85-28145

Fax: +49 9131 302941

[krasimir.zhelev@rrze.uni-erlangen.de](mailto:krasimir.zhelev@rrze.uni-erlangen.de)

<http://www.rrze.uni-erlangen.de>

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)



Regionales  
Rechenzentrum  
Erlangen

Der IT-Dienstleister der FAU