QuickMig

Semi-automatic Schema Matching for Data Migration

SYSTEMATIC THOUGHT LEADERSHIP FOR INNOVATIVE BUSINESS

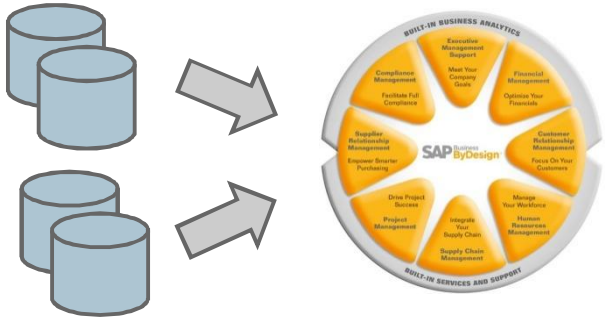
Christian Drumm,	SAP AG
Matthias Schmitt,	SAP AG
Hong-Hai Do,	SAP AG
Erhard Rahm,	University Leipzig

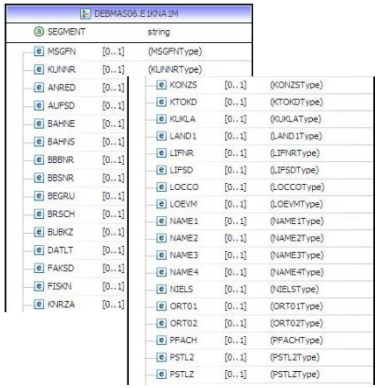
01.10.2007

QuickMig: Overview

Problem Description

- Data migration from unknown source to well-known target system
- Data export schemas of legacy systems not designed for interoperability
- Development of mappings requires significant effort





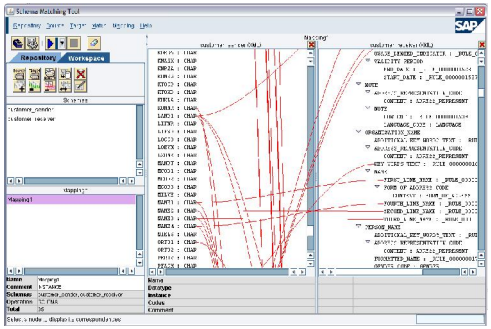
Automatic Schema Matching

- COMA++
  - Used to evaluate schema based matching approaches
  - Enables flexible combination of different matching approaches

➔ Schema-based matching approaches not applicable to data migration use case

Solution

- Development of new matching algorithms
- Development of a new methodology
- Combination of new approaches with existing COMA++ functionality
- Prototype implementation including evaluation of new approach



SAP RESEARCH

Agenda

1. **QuickMig**
    - 1.1. Data Migration
    - 1.2. New Concepts
    - 1.3. Evaluation
2. Open Research Questions

Data Migration: Scenario

Scenario

- Data migration for „SAP Business ByDesign“
  - SAP's comprehensive new mid market solution
  - Financials, Customer Relationship Management, Human Resources Management, Supply Chain Management, Project Management, …
- Volume business
- Legacy data migration within one week

Challenges

- Customers lack the expertise to perform data migration
- Customers cannot afford expensive data migration projects
- High diversity of possible source systems
  - 3rd party systems
  - Custom solutions

➔ **Migration of any (unknown) source system to well-known targets system**

Data Migration: Opportunities – Domain Knowledge

Limited value of source schema information

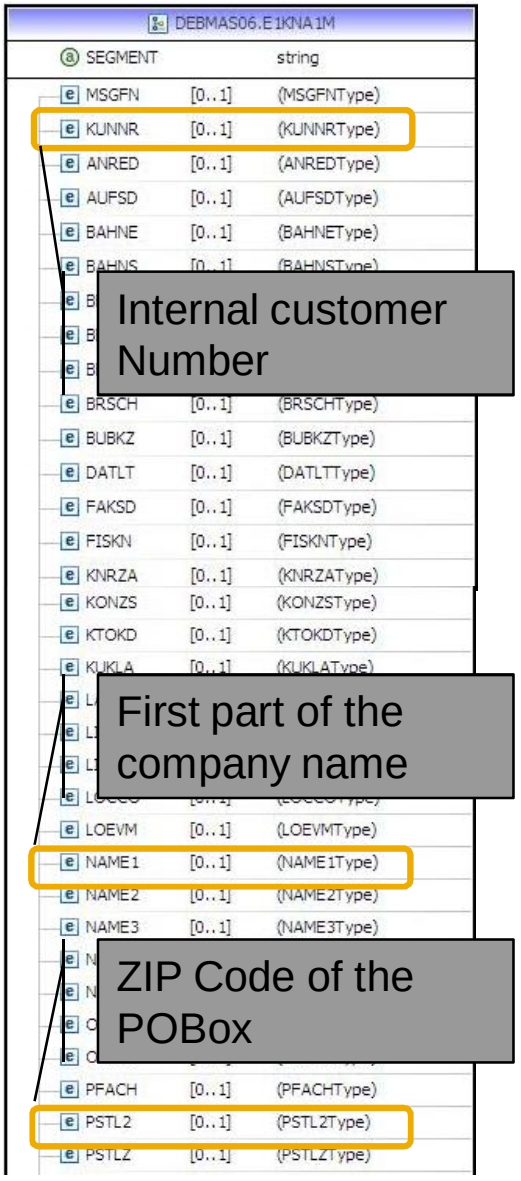
- Data export schemas of legacy systems are not designed for interoperability
- Technical names, abbreviations, proprietary structures, codes, flat structures, …
- Low quality of meta-data (e.g. element documentation)

Availability of domain knowledge

- Software vendor has detailed knowledge about business capabilities and import interfaces of target system
- Customer has detailed knowledge about business capabilities of source system

➔ **Utilize target system knowledge**

➔ **Leverage business knowledge of source system**



Data Migration: Opportunities – Scope

Scope of data migration projects

- Rich target system capabilities which can be easily configured to customers needs
- Data Migration highly dependent on source system capabilities

Accumulated mapping knowledge

- Several mapping tasks in one data migration project
  e.g. mapping of customer data, supplier data, purchase orders, …
- Schemas usually contain similar elements and sub-structures
  e.g. address data, bank data, currency code, …

➔ **Flexibly reduce migration scope**

➔ **Reuse previous mapping results**

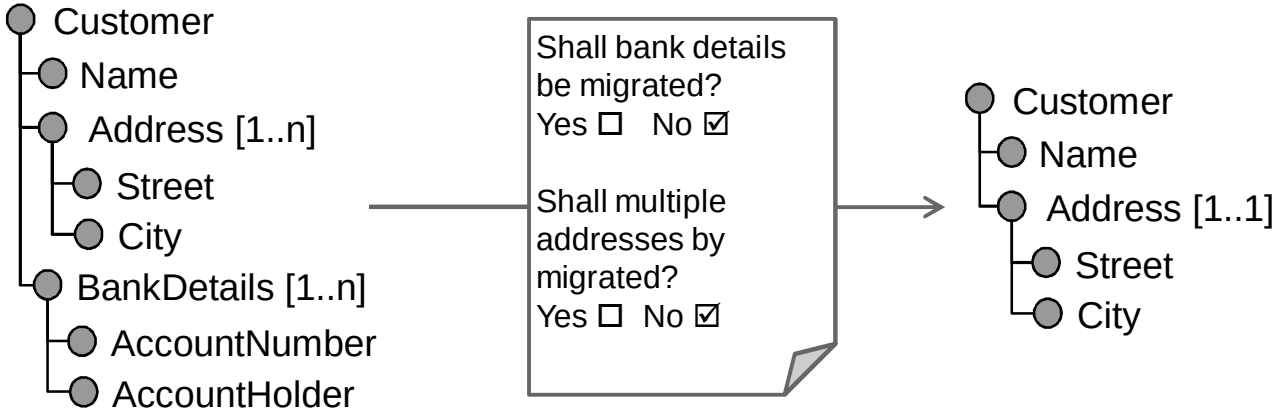
New Concepts: Schema Reduction

Goal

- Reduce complexity of matching task
- Simplify verification of proposed mapping

Approach

- Electronic questionnaire to exploit source system capabilities and customer specific scope
- Each question targets a specific capability of the target system
  - Irrelevant parts of target schema can be removed
  - Additional simplifications (e.g. cardinality, time dependency, …)
- Mappings between original schema and reduced schema can be derived automatically



New Concepts: Sample Data

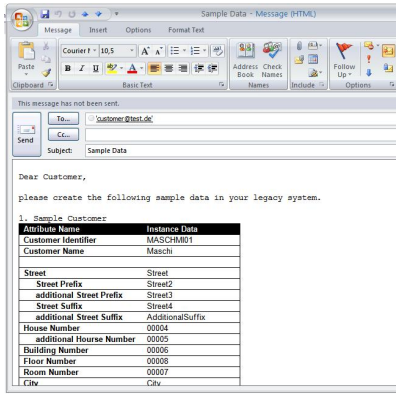
Goal

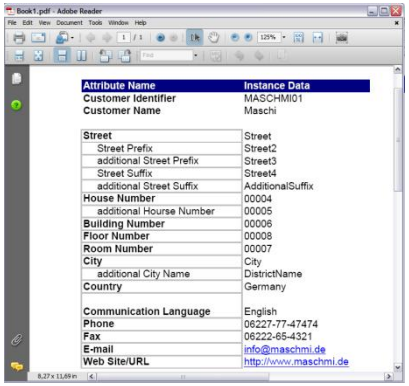
- Use instance data for schema matching
- Achieve high precision value (=1.0)

Approach

- Pre-deliver specially prepared instance data for target structure
  - For elements with explicit business content only (no technical elements, no codes, …)
  - Highly selective content for every element (no identical values, no substrings, …)
- Enforce identical instance in source system
  - Sample instance is provided in business terms (PDF, Excel, Email, …)
  - Customer can easily create sample instance via standard business user interface

	A	B
1	**Attribute Name**	**Instance Data**
2	**Customer Identifier**	MASCHMI01
3	**Customer Name**	Maschi
4		
5	**Street**	Street
6	Street Prefix	Street2
7	additional Street Prefix	Street3
8	Street Suffix	Street4
9	additional Street Suffix	AdditionalSuffix
10	**House Number**	00004
11	additional Hourse Number	00005
12	**Building Number**	00006
13	**Floor Number**	00008
14	**Room Number**	00007
15	**City**	City
16	additional City Name	DistrictName
17	**Country**	Germany
18	**Communication Language**	English
19	**Phone**	06227-77-47474
20	**Fax**	06222-65-4321
21	**E-mail**	info@maschmi.de
22	**Web Site/URL**	http://www.maschmi.de





New Concepts: Domain Knowledge

Goal

- Utilize detailed target system knowledge
- Achieve higher recall value

Approach

- Model domain knowledge
  - Alternative information (about data formats and common standards)
    - Different date and time formats
    - Phone Number, Street and House Number
  - Additional information (about target system)
    - Code lists including texts describing the code value
    - Default values for required fields (context dependent)
    - Query attributes for resolution of foreign key associations
- Provide additional instance based matchers exploiting domain knowledge

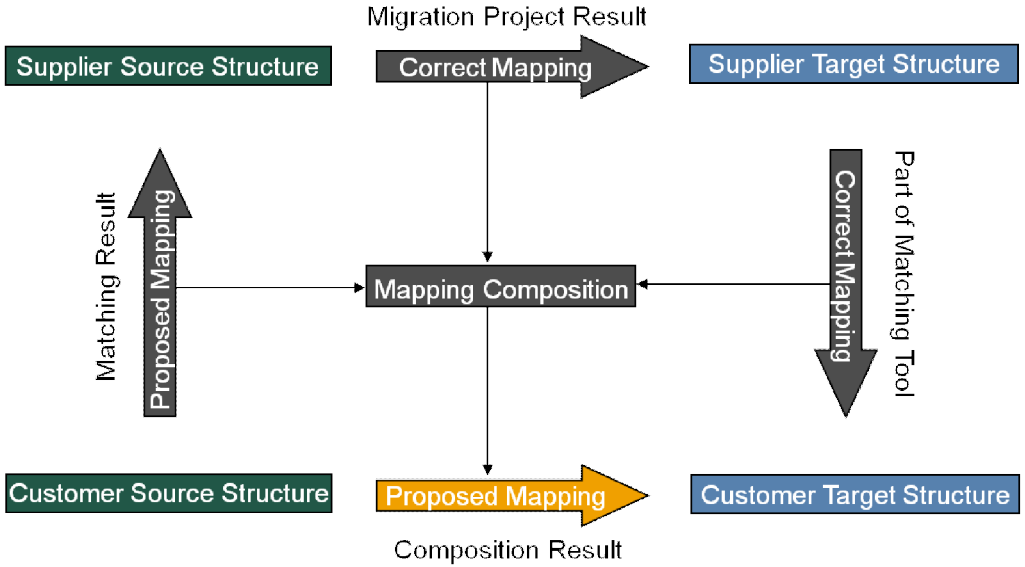
New Concepts: Reuse

Goal

- Reuse previous mapping results within one data migration project
- Achieve higher recall value

Approach

- Provide mapping between similar structures of the target system
- Match corresponding source structures
- Compose required mapping by reusing previously developed correct mapping



New Concepts: Mapping Categories

Goal

- Determine executable mapping code (beyond simple “Move”)

Approach

- Develop a list of semantic mapping categories covering all required transformations
- Extend matchers to enable determination of mapping categories by exploiting sample data and domain knowledge
- Extend “Reuse” matcher to support composition of mapping categories
- Derive mapping code in runtime system
  - Generation of mapping code (“Move”, “Concatenate”, …)
  - Generation of code templates (“Value Mapping”, “Internal ID”, …)
  - Manual implementation (“Complex”)

Mapping Categories
Create Instance
Key Mapping
Internal ID
Look Up
Move
Value Mapping
Default Value
Split
Concatenate
Complex

Evaluation: Schema Reduction

Evaluation

- Schemas used for evaluation*

Scenario	# of Elements
Target Schema	4639
SAP R/3 4.0	953
SAP ERP	2150
SAP B1	480

- Target schema after reduction

Scenario	# of Elements
Target Schema	4639
SAP R/3 4.0	645
SAP ERP	612
SAP B1	639

➔ **Resulting target schemas differ largely depending on source system**

➔ **Significant reduction of target schema possible**

* Evaluation data sets are available at: https://www.sdn.sap.com/irj/sdn/weblogs?blog=/pub/wlg/7008

Evaluation: Sample Data

Evaluation

- Purely schema based matching

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	0.00	0.00	0.00
SAP ERP	0.40	0.04	0.08
SAP B1	0.31	0.18	0.23
Average	0.23	0.07	0.10

- Sample data based matching

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	1	0.27	0.43
SAP ERP	1	0.38	0.55
SAP B1	1	0.58	0.70
Average	1	0.41	0.56

➔ **Schema-based matching approaches not applicable**

➔ **Sample data matcher achieves precision of 1**

➔ **Improvement of recall required**

Evaluation: Domain Knowledge

Evaluation

- Sample data based matching

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	1	0.27	0.43
SAP ERP	1	0.38	0.55
SAP B1	1	0.58	0.70
Average	1	0.41	0.56

- Combination of domain knowledge and sample data

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	1	0.50	0.66
SAP ERP	1	0.49	0.65
SAP B1	1	0.65	0.79
Average	1	0.55	0.70

➔ **Utilization of domain knowledge significantly improves recall**

➔ **Matcher still achieves precision of 1**

Evaluation: Reuse

Evaluation

- Reuse matcher

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	0.80	0.50	0.61
SAP ERP	0.81	0.43	0.56
SAP B1	1	0.80	0.89
Average	0.87	0.58	0.69

- Combination of reuse, sample data and domain knowledge

Scenario	Prec.	Recall	F-Meas.
SAP R/3 4.0	0.99	0.67	0.80
SAP ERP	0.99	0.70	0.82
SAP B1	1	0.80	0.89
Average	0.99	0.72	0.84

➔ **Reuse matcher achieves mediocre precision and recall values**

➔ **Reuse matcher and sample data determine complementary matches**

➔ **Combination of all approaches delivers very good results**

Evaluation: Mapping Categories

Evaluation

- 97% of mapping categories identified correctly
- Automatic code generation for more than 40% of the matches possible („Move“ and „Split“)

Mapping Categories	Occurrence
Create Instance	0.07
Key Mapping	0.02
Internal ID	0.02
Look Up	0.13
Move	0.36
Value Mapping	0.14
Default Value	0.02
Split	0.07
Concatenate	0.00
Complex	0.15

➔ **Significant reduction of manual implementation effort possible**

Agenda

1. SAP Research
2. QuickMig
3. **Open Research Questions**

Open Research Questions

1. **How can mapping code be generated automatically based on mapping categories?**
   - Obvious in case of „Move“
   - What about more complex categories
   - Templates my allready be very helpful
2. **How can (useful) user input be aquired to support the automatic mapping process?**
   - Sample data just a first step
   - Important that user input is aquired non-intrusively
   - Where is the sweet-spot?
3. **How to handle codes and code lists**
   - Sample data not helpful due to large number of similar values
   - Unknown codes and code lists

Thank you!

QuickMig: Resources

COMA++

- http://dbs.uni-leipzig.de/de/Research/coma.html

QuickMig

- CIKM 2007 Paper: http://dbs.uni-leipzig.de/file/QuickMig-cikm07.pdf
- Evaluation data sets: https://www.sdn.sap.com/irj/sdn/weblogs?blog=/pub/wlg/7008

Contact Details

- Christian Drumm christian.drumm@sap.com
- Hong-Hai Do hong-hai.do@sap.com

Copyright 2007 SAP AG
All rights reserved