

QuickMig

Semi-automatic Schema Matching for Data Migration

SYSTEMATIC THOUGHT LEADERSHIP FOR INNOVATIVE BUSINESS



Christian Drumm,
Matthias Schmitt,
Hong-Hai Do,
Erhard Rahm,

SAP AG
SAP AG
SAP AG
University Leipzig

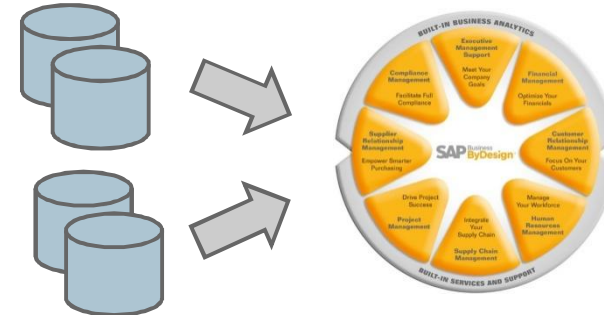
01.10.2007

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



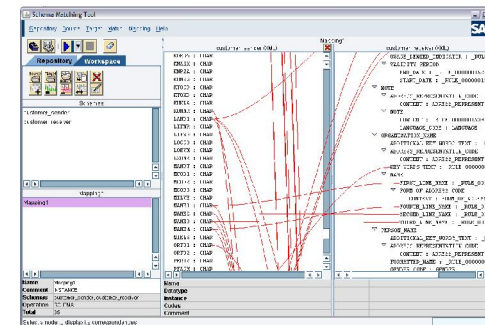
SEGMENT	string
[E] MSGFN [0..1] (MSGFNType)	[E] KONZS [0..1] (KONZSType)
[E] KUNNR [0..1] (KUNNRType)	[E] KTOKD [0..1] (KTOKDType)
[E] ANRED [0..1] (ANREDType)	[E] KUKLA [0..1] (KUKLAType)
[E] ALFSD [0..1] (ALFSDType)	[E] LAND1 [0..1] (LAND1Type)
[E] BAHNE [0..1] (BAHNEType)	[E] LIPFR [0..1] (LIPFRType)
[E] BAHVS [0..1] (BAHVSType)	[E] LIFS0 [0..1] (LIFS0Type)
[E] BBSNR [0..1] (BBSNRType)	[E] LOCCO [0..1] (LOCCOType)
[E] BEGRU [0..1] (BEGRUType)	[E] LOEVM [0..1] (LOEVMTType)
[E] BRSCH [0..1] (BRSCHType)	[E] NAME1 [0..1] (NAME1Type)
[E] SUBKZ [0..1] (SUBKZType)	[E] NAME2 [0..1] (NAME2Type)
[E] DATLT [0..1] (DATLTType)	[E] NAME3 [0..1] (NAME3Type)
[E] PAKSD [0..1] (PAKSDType)	[E] NAME4 [0..1] (NAME4Type)
[E] PISKN [0..1] (PISKNType)	[E] NIELS [0..1] (NIELSType)
[E] KNRZA [0..1] (KNRZAType)	[E] ORT01 [0..1] (ORT01Type)
	[E] ORT02 [0..1] (ORT02Type)
	[E] PFACH [0..1] (PFACHType)
	[E] PSTL2 [0..1] (PSTL2Type)
	[E] PSTLZ [0..1] (PSTLZType)

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



Agenda



SAP

1. QuickMig

- 1.1. Data Migration
- 1.2. New Concepts
- 1.3. Evaluation

2. Open Research Questions

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

Element Name	Cardinality	Data Type
MSGFN	[0..1]	(MSGFNType)
KUNNR	[0..1]	(KUNNRType)
ANRED	[0..1]	(ANREDType)
AUFSD	[0..1]	(AUFSDType)
BAHNE	[0..1]	(BAHNType)
BAHNS	[0..1]	(BAHNType)
BRSCH	[0..1]	(BRSCHType)
BUBKZ	[0..1]	(BUBKZType)
DATLT	[0..1]	(DATLTType)
FAKSD	[0..1]	(FAKSDType)
FISKN	[0..1]	(FISKNType)
KNRZA	[0..1]	(KNRZAType)
KONZS	[0..1]	(KONZSType)
KTOKD	[0..1]	(KTOKDType)
KUKLA	[0..1]	(KUKLAType)
NAME1	[0..1]	(NAME1Type)
NAME2	[0..1]	(NAME2Type)
NAME3	[0..1]	(NAME3Type)
PFACH	[0..1]	(PFACHType)
PSTL2	[0..1]	(PSTL2Type)
PSTLZ	[0..1]	(PSTLZType)



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

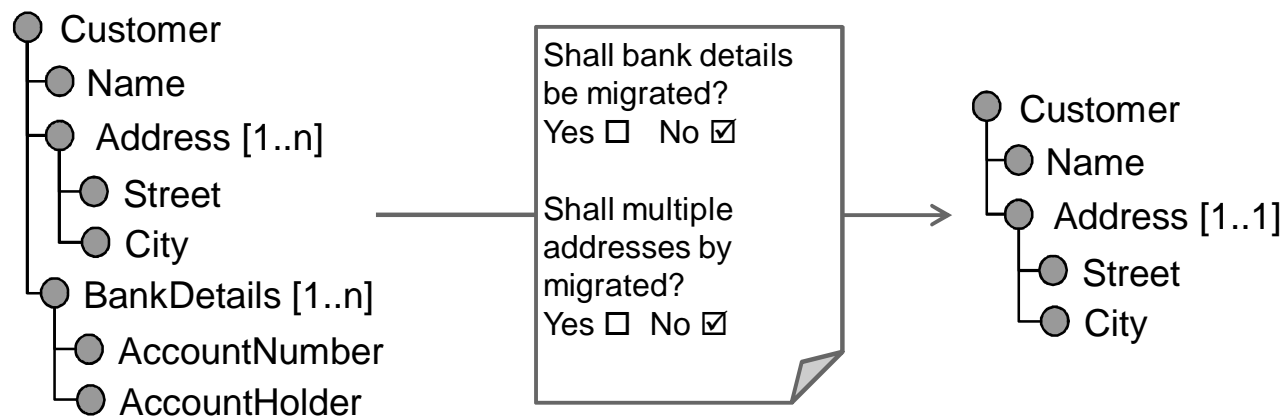
➔ Reuse previous mapping results

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



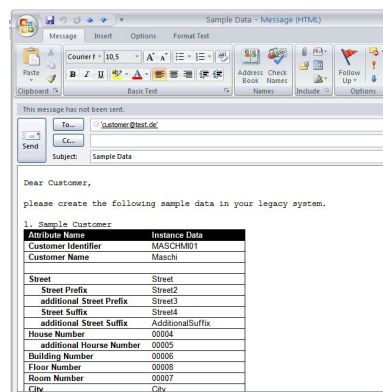
Goal

- 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

Attribute Name	Instance Data
Customer Identifier	MASCHMI01
Customer Name	Maschi
Street	Street
Street Prefix	Street2
additional Street Prefix	Street3
Street Suffix	Street4
additional Street Suffix	AdditionalSuffix
House Number	00004
additional House Number	00005
Building Number	00006
Floor Number	00008
Room Number	00007
City	City
additional City Name	DistrictName
Country	Germany
Communication Language	English
Phone	06227-77-47474
Fax	06222-65-4321
E-mail	info@maschmi.de
Web Site/URL	http://www.maschmi.de



Attribute Name	Instance Data
Customer Identifier	MASCHMI01
Customer Name	Maschi
Street	Street
Street Prefix	Street2
additional Street Prefix	Street3
Street Suffix	Street4
additional Street Suffix	AdditionalSuffix
House Number	00004
additional House Number	00005
Building Number	00006
Floor Number	00008
Room Number	00007
City	City
additional City Name	DistrictName
Country	Germany
Communication Language	English
Phone	06227-77-47474
Fax	06222-65-4321
E-mail	info@maschmi.de
Web Site/URL	http://www.maschmi.de

Goal

- Utilize detailed target system knowledge
- Achieve higher recall value

Approach

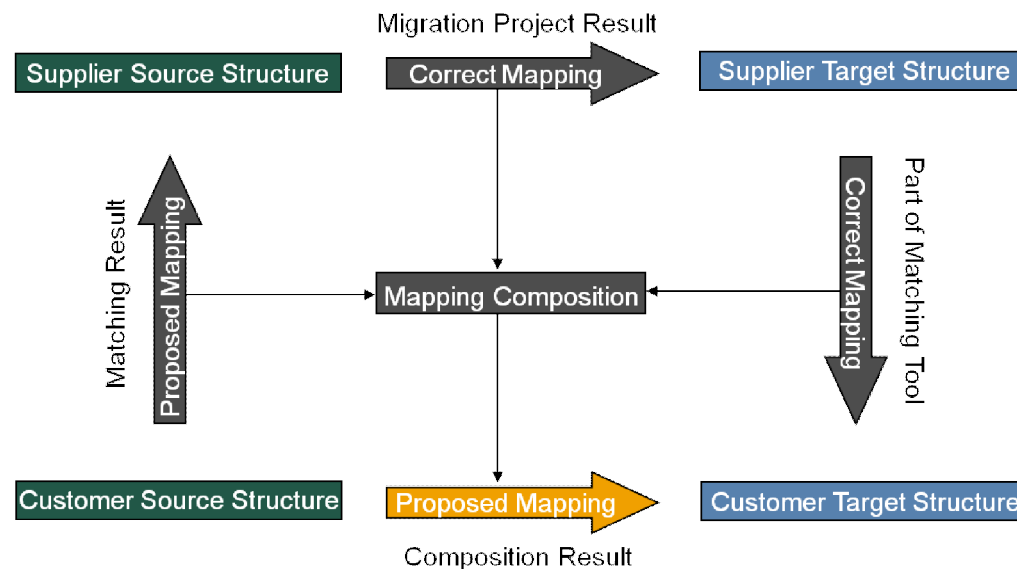
- Model domain knowledge
 - Alternative information (about data formats and common standards)
 - Different date and time formats
 - Phone Number, Street and House Number
 - Additional information (about target system)
 - Code lists including texts describing the code value
 - Default values for required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

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



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

■ 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

■ 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

- 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 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**

1. **How can mapping code be generated automatically based on mapping categories?**
 - Obvious in case of „Move“
 - What about more complex categories
 - Templates may already be very helpful

2. **How can (useful) user input be acquired to support the automatic mapping process?**
 - Sample data just a first step
 - Important that user input is acquired **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!



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



No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

SAP, R/3, mySAP, mySAP.com, xApps, xApp, SAP NetWeaver, Duet, Business ByDesign, ByDesign, PartnerEdge and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world. All other product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. This document is a preliminary version and not subject to your license agreement or any other agreement with SAP. This document contains only intended strategies, developments, and functionalities of the SAP® product and is not intended to be binding upon SAP to any particular course of business, product strategy, and/or development. SAP assumes no responsibility for errors or omissions in this document. SAP does not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

SAP shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The statutory liability for personal injury and defective products is not affected. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages

Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG nicht gestattet. In dieser Publikation enthaltene Informationen können ohne vorherige Ankündigung geändert werden.

Einige von der SAP AG und deren Vertriebspartnern vertriebene Softwareprodukte können Softwarekomponenten umfassen, die Eigentum anderer Softwarehersteller sind.

SAP, R/3, mySAP, mySAP.com, xApps, xApp, SAP NetWeaver, Duet, Business ByDesign, ByDesign, PartnerEdge und andere in diesem Dokument erwähnte SAP-Produkte und Services sowie die dazugehörigen Logos sind Marken oder eingetragene Marken der SAP AG in Deutschland und in mehreren anderen Ländern weltweit. Alle anderen in diesem Dokument erwähnten Namen von Produkten und Services sowie die damit verbundenen Firmenlogos sind Marken der jeweiligen Unternehmen. Die Angaben im Text sind unverbindlich und dienen lediglich zu Informationszwecken. Produkte können länderspezifische Unterschiede aufweisen.

Die in diesem Dokument enthaltenen Informationen sind Eigentum von SAP. Dieses Dokument ist eine Vorabversion und unterliegt nicht Ihrer Lizenzvereinbarung oder einer anderen Vereinbarung mit SAP. Dieses Dokument enthält nur vorgesehene Strategien, Entwicklungen und Funktionen des SAP®-Produkts und ist für SAP nicht bindend, einen bestimmten Geschäftsweg, eine Produktstrategie bzw. -entwicklung einzuschlagen. SAP übernimmt keine Verantwortung für Fehler oder Auslassungen in diesen Materialien. SAP garantiert nicht die Richtigkeit oder Vollständigkeit der Informationen, Texte, Grafiken, Links oder anderer in diesen Materialien enthaltenen Elemente. Diese Publikation wird ohne jegliche Gewähr, weder ausdrücklich noch stillschweigend, bereitgestellt. Dies gilt u. a., aber nicht ausschließlich, hinsichtlich der Gewährleistung der Marktgängigkeit und der Eignung für einen bestimmten Zweck sowie für die Gewährleistung der Nichtverletzung geltenden Rechts.

SAP übernimmt keine Haftung für Schäden jeglicher Art, einschließlich und ohne Einschränkung für direkte, spezielle, indirekte oder Folgeschäden im Zusammenhang mit der Verwendung dieser Unterlagen. Diese Einschränkung gilt nicht bei Vorsatz oder grober Fahrlässigkeit.

Die gesetzliche Haftung bei Personenschäden oder die Produkthaftung bleibt unberührt. Die Informationen, auf die Sie möglicherweise über die in diesem Material enthaltenen Hotlinks zugreifen, unterliegen nicht dem Einfluss von SAP, und SAP unterstützt nicht die Nutzung von Internetseiten Dritter durch Sie und gibt keinerlei Gewährleistungen oder Zusagen über Internetseiten Dritter ab.

Alle Rechte vorbehalten.