A Validatable Legacy Database Migration Using ORM

Explanation of a legacy database migration method using ORM and demonstration of its application in a case

Tjeerd Moes, TNO
Jan Pieter Wijbenga, TNO
Herman Balsters, University of Groningen
George Huitema, University of Groningen
Legacy database migration

▷ What is a legacy information system?
  ◇ any system that resists modification and evolution

▷ Why is it a problem?
  ◇ system can’t adapt to changing business process
  ◇ hard to maintain and extend
  ◇ insufficient performance
  ◇ risk due to chance of failure (with serious impact)

▷ Why migrate the database?
  ◇ preserve valuable data
  ◇ legacy system can be phased out eventually
Types of migration projects

- Greenfield target
- Single source
- Fixed target
- Multiple sources
From legacy source to fixed target

Hard problem

No documentation
Semantics unknown
Missing primary keys
Dirty data

Fixed target
Which data?

Missing referential integrity constraints
Differences in schema design

Which data?
State of the art

- Focus of literature is on technical aspects

- Black-box approach:
  - lack of focus on making migration decisions explicit
  - hence, hard to validate these decisions

- Little focus on source system understanding
- Little focus recovery and validation of semantics

- No method that applies fact oriented modeling in legacy database migration with fixed target

- Not much practical experience with application of conceptual data modeling in legacy database migration, UML and ER only
Goal

› Transfer the data from the source system to the fixed target while preserving the semantics in order to provide correct management information

› Perform the migration in a repeatable fashion
  › using fact-oriented modeling
  › applying ORM Logic-based English (OLE)
Requirements for the method

Our method should

- facilitate communication between stakeholders
- enable extraction and formalization of semantics
- apply a white-box approach
Case: online survey systems

2 survey systems about customer satisfaction used by a Dutch sales company

KBM

LimeSurvey
Old and new survey systems

<table>
<thead>
<tr>
<th></th>
<th>zeer</th>
<th>oneens</th>
<th>oneens</th>
<th>eens</th>
<th>zeer</th>
<th>eens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. De medewerk(st)er aan de telefoon was</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vriendelijk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b. De medewerk(st)er aan de telefoon heeft me</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>goed geholpen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c. De bezorger van mijn bestelling was vriendelijk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d. De algemene indruk die ik van de bezorger kreeg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>was goed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Toelichting of opmerkingen over het contact medewerkers met u hebben opgenomen:
Old and new survey systems

---

**Order- en leverproces**

0% 0% 100%

**Onderdeel 3**

**Communicatie**

5

Onderstaande vragen gaan over het contact dat medewerkers met u hebben opgenomen. Kies "n.v.t." als u over een onderwerp geen uitspraak kunt of wilt doen, bijv. als uw bestelling niet aan de deur is afgeleverd.

<table>
<thead>
<tr>
<th></th>
<th>zeer oneens</th>
<th>oneens</th>
<th>eens</th>
<th>zeer eens</th>
<th>n.v.t.</th>
</tr>
</thead>
<tbody>
<tr>
<td>De medewerk(a)ter aan de telefoon was vriendelijk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De medewerk(a)ter aan de telefoon heeft me goed geholpen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De bezorger van mijn bestelling was vriendelijk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De algemene indruk die ik van de bezorger kreeg was goed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Toelichting of opmerkingen over het contact medewerkers met u hebben opgenomen:

---

7 Heeft u zelf contact moeten opnemen over uw bestelling?

Kies één van de volgende antwoorden:

- Ja, ik heb contact opgenomen n.a.v. mijn bestelling
- Nee, ik heb geen contact opgenomen.

Bijvoorbeeld per e-mail of telefonisch
Example of schema difference

Universe of discourse

Source

<table>
<thead>
<tr>
<th>kbm_Consumer</th>
<th>kbm_Con-Char-Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>id</td>
</tr>
<tr>
<td></td>
<td>mailaddress</td>
</tr>
<tr>
<td></td>
<td>extracharacteristic1</td>
</tr>
<tr>
<td></td>
<td>extracharacteristic2</td>
</tr>
</tbody>
</table>

Target

<table>
<thead>
<tr>
<th>ls_tokens_53198</th>
<th>ls_tokens_71266</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>tid</td>
</tr>
<tr>
<td></td>
<td>email</td>
</tr>
<tr>
<td></td>
<td>attribute1</td>
</tr>
<tr>
<td></td>
<td>attribute2</td>
</tr>
<tr>
<td></td>
<td>attribute3</td>
</tr>
</tbody>
</table>
We introduce the optimized model

Model of the universe of discourse

- Facilitates communication and validation
- Contains relevant selection of data
- Provides a deliverable

Created using:
- ORM 2’s conceptual schema design procedure
- NORMA tool
Extracting semantics

- No documentation
- Semantics unknown
- Missing primary keys
- Dirty data
- Missing referential integrity constraints
Extracting semantics

UNOPTIMIZED SOURCE ORM MODEL

UNOPTIMIZED TARGET ORM MODEL

Basic transformations
Method: Mapping

Mapping ORM models requires a conceptual query language. (OLE)
Results

› Demo
› We will be showing the running example from the source to the optimized to the target.
Onderdeel
Stap 5 van 19

De volgende vragen gaan over uw plannen en beweegredenen omtrent IPv6

1. Bent u op de hoogte van de komst van IPv6?
   - Ja
   - Nee

Volgende
Source Application KBM screenshots

Order- en leverproces

Enquêtegroep:
- groep-Vragenstructuur
- groep-Responsinfo
- groep-Grafieken

Grafieken en ruwe data worden bij de knop 'Genereer grafieken en ruwe data' alleen gegenereerd als ze nog niet bestonden. Dit dus het kan enige tijd duren!


Enquête:
- enq-Vragenstructuur
- eng-Responsinfo
- eng-Valideer
- enq-Grafieken
- enq-RuweData
- echtAlleRuweData
- Test
Source Application KBM screenshots

Number of **consumer characteristics**
(The number of consumer columns (excluding email address) in the inviter sheet and in the raw data. Regular is ‘4’.)

Names of **consumer-characteristics**. Before you change these setting: **warn the system administrator!**
If a text is provided, this text will be used as name of the first consumer-characteristics.
If no text is provided, the name from the Client table will be used (only if the number of characteristics is 4 or less) (...)

<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>email address</td>
</tr>
<tr>
<td>2</td>
<td>Sales channel</td>
</tr>
<tr>
<td>3</td>
<td>Brand code</td>
</tr>
<tr>
<td>4</td>
<td>External order ID</td>
</tr>
<tr>
<td>5</td>
<td>V-code</td>
</tr>
<tr>
<td>6</td>
<td>Order type description</td>
</tr>
<tr>
<td>7</td>
<td>Order subtype description</td>
</tr>
<tr>
<td>8</td>
<td>Logistics partner</td>
</tr>
</tbody>
</table>
### Source: KBM
Database schema

<table>
<thead>
<tr>
<th>consumer</th>
<th>consumercharacteristic</th>
<th>providedanswer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>id</td>
<td></td>
</tr>
<tr>
<td></td>
<td>consumergroupid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mailaddress</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extracharacteristic1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extracharacteristic2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extracharacteristic3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extracharacteristic4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>monitorstatus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>monitorstartdate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>monitornextdate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>monitorcategory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>monitorphonenumnumber</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shipmentid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>id</td>
<td></td>
</tr>
<tr>
<td></td>
<td>consumerid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ordernr</td>
<td></td>
</tr>
<tr>
<td></td>
<td>text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>optout</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mailaddress</td>
<td></td>
</tr>
<tr>
<td></td>
<td>askedquestionid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>answeroption</td>
<td></td>
</tr>
<tr>
<td></td>
<td>number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>freetext</td>
<td></td>
</tr>
<tr>
<td></td>
<td>freetext2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>responseid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>l2</td>
<td></td>
</tr>
</tbody>
</table>
Target: LS
Database schema
**Target: LS**

**Database schema**

<table>
<thead>
<tr>
<th>survey_53198</th>
<th>tokens_53198</th>
<th>surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>id</td>
<td>PK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>submitdate</td>
<td></td>
<td>firstname</td>
</tr>
<tr>
<td>lastpage</td>
<td></td>
<td>lastname</td>
</tr>
<tr>
<td>startlanguage</td>
<td></td>
<td>email</td>
</tr>
</tbody>
</table>
Experienced difficulties

- How to find out semantics?
- How to make a mapping?
- How to deal with implicit constraints?
Step 1: Optimized Conceptual Model
Step 1: Optimized Conceptual Model
Step 2: Unoptimized Source Model
Step 2: Unoptimized Source Model
Step 2: Unoptimized Source Model
Step 2: Unoptimized Source Model
Step 3: Mapping

Source schema

Kbm_Consumer
- Columns
  - PK: id: int
  - kbm_mailaddress: varchar(255)
  - FK1: zendingid: int
  - extraCharacteristic1: varchar(255)
  - extraCharacteristic2: varchar(255)

Kbm_ConsumerCharacteristic
- Columns
  - PK: id: int
  - U1, FK1: consumentid: int
  - U1: orderNr: int
  - text: varchar(255)

Optimized schema

M_Consumer
- Columns
  - PK: m_Consumer: int
  - FK1: m_MailAddress: varchar(128)
  - FK2: survey_id: int
  - datetime: timestamp

M_Characteristic
- Columns
  - PK, FK1: consumer_id: int
  - PK, FK2: name: varchar(128)
  - value: varchar(128)

**OLE**

for each m_Characteristic and Value, that m_Characteristic has that Value iff there is some kbm_ConsumerCharacteristic, such that

m_Characteristic.consumer_id = kbm_ConsumerCharacteristic.consumerid

and

(Value = the text of kbm_ConsumerCharacteristic)

or

Value = the extraCharacteristic1 of kbm_Consumer

or

Value = the extraCharacteristic2 of kbm_Consumer)
Step 3: Mapping

**OLE**

for each m_Characteristic and Value, that m_Characteristic has that Value iff there is some kbm_ConsumerCharacteristic, such that m_Characteristic.consumer_id = kbm_ConsumerCharacteristic.consumerid and (Value = the text of kbm_ConsumerCharacteristic) or Value = the extraCharacteristic1 of kbm_Consumer or Value = the extraCharacteristic2 of kbm_Consumer)

**SQL VIEW**

CREATE VIEW m_Characteristic AS
SELECT kbm_CCh.consumerid AS consumer_id, kbm_CCh.text AS value, [..] FROM kbm_ConsumerCharacteristic kbm_CCh, [..]
UNION SELECT kbm_C.id AS consumer_id, kbm_C.extraCharacteristic1 AS value FROM kbm_Consumer kbm_C, [..].
Implementation – Replication service
Implementation – VIEW definition

```
FROM (SELECT eg.id AS AS_enquetegroepid, MAX(e.id) AS maxenquetegroepid
     FROM kbm dbo.enquetegroep AS eg ON e.enquetegroepid = eg.id
     GROUP BY eg.id) AS egm enquetegroepid
     kbm dbo.enquetegroepmigratie AS egmig ON egm.enquetegroepid = egmig.id INNER JOIN
     kbm dbo.pagina AS p ON egm.maxenquetegroepid = p.enquetegroepid INNER JOIN
     kbm dbo.voorwaarde/raag AS vw ON vw.voorwaardeid = v.id AND vw.paginaid = p.id

<table>
<thead>
<tr>
<th>Column</th>
<th>Alias</th>
<th>Table</th>
<th>Output</th>
<th>Sort Type</th>
<th>Sort Order</th>
<th>Filter</th>
<th>Or...</th>
<th>Or...</th>
<th>Or...</th>
</tr>
</thead>
<tbody>
<tr>
<td>enquetegroepid</td>
<td>m_enquete</td>
<td>egme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>volgrn</td>
<td>m_paginanr</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vraagnr</td>
<td></td>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vraagcode</td>
<td>referenteCode</td>
<td>v</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tekst</td>
<td>vraagTekst</td>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>antwoordpres...</td>
<td>vraagType</td>
<td>v</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>m_enquete</th>
<th>m_paginanr</th>
<th>vraagnr</th>
<th>referenteCode</th>
<th>vraagTekst</th>
<th>vraagType</th>
<th>helpTekst</th>
<th>isVerplicht</th>
</tr>
</thead>
<tbody>
<tr>
<td>3001</td>
<td>2</td>
<td>18</td>
<td>Fol_1_8</td>
<td>Bent u tevreden over de dienst ...</td>
<td>2</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>3001</td>
<td>2</td>
<td>19</td>
<td>Fol_1_9</td>
<td>Bent u tevreden over de dienst ...</td>
<td>2</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>3001</td>
<td>2</td>
<td>20</td>
<td>Fol_1_A</td>
<td>Bent u tevreden over de dienst ...</td>
<td>2</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>
```
Experiences

- Improved knowledge sharing and validation by introducing reengineered ORM models
- Divide mapping problem into smaller pieces using the intermediate, optimized ORM model
- Increased openness using conceptual query language
- A controllable and systematic process
- Intermediate deliverables
Discussion

Requirements (again)
› facilitate communication between stakeholders
› enable extraction and formalization of semantics
› apply a white-box approach

› Extracting and cleaning was facilitated
› ORM and use of the unoptimized models generated knowledge of the source and target
› Using formal procedures makes the process repeatable
Discussion

- Two mappings
- The role of the unoptimized model
- Scalability
Future research directions

› Advancement of a natural language-like conceptual query language as FORML, OLE and CQL

› How to address a legacy landscape (interrelated legacy systems)