Chapter XXI Database Reverse Engineering

Jean-Luc Hainaut University of Namur, Belgium

> Jean Henrard REVER s.a., Belgium

Didier Roland

REVER s.a., Belgium

Jean-Marc Hick REVER s.a., Belgium

Vincent Englebert University of Namur, Belgium

INTRODUCTION

Database reverse engineering consists of recovering the abstract descriptions of files and databases of legacy information systems. A legacy information system can be defined as a "data-intensive application, such as [a] business system based on hundreds or thousands of data files (or tables), that significantly resists modifications and changes" (Brodie & Stonebraker, 1995). The objective of database reverse engineering is to recover the logical and conceptual descriptions, or schemas, of the permanent data of a legacy information system, that is, its database, be it implemented as a set of files or through an actual database management system. The logical schema is the technology-dependent (e.g., relational) description of the database structures while the conceptual schema is the abstract, technology-independent description of their semantics.

Database reverse engineering often is the first steps of broader engineering projects. Indeed, rebuilding the precise documentation of a legacy database is an absolute prerequisite before migrating, reengineering, maintaining or extending it, or merging it with other databases.

The current commercial offering in CASE tools poorly supports database reverse engineering. Generally, it reduces to the straightforward derivation of a conceptual schema such as that of Figure 1 from the following DDL code.

T. 1			C 1				
HIGHING I	$\Delta nawa$	1/10141	nt d	ata	VONOVCO	ongino	ovino
Γ I Σ U I C I L	Anuive	VIEW	u	uiu	IEVEISE	eneme	erine
			-,				

```
create table CUSTOMER (
CNUM decimal(10) not null,
CNAME varchar(60) not null,
CADDRESS varchar(100) not null,
primary key (CNUM))
cate table ORDER (
    ONUM decimal(12) not null,
    SENDER decimal(10) not null,
    ODATE date not null,
    primary key (ONUM),
    foreign key (CNUM) references CUSTOMER))
```



Unfortunately, actual database reverse engineering often is closer to deriving the conceptual schema of Figure 2 from the following sections of COBOL code, using meaningless names that do not declare compound fields or foreign keys.

Getting such a result obviously requires additional sources of information, which may prove more difficult to analyze than mere DDL statements. Untranslated (implicit) data structures and constraints, empirical implementation approaches and techniques, optimization constructs, ill-designed schemas, and, above all, the lack of up-to-date documentation are some of the difficulties that the analysts will face when trying to understand existing databases. The goal of this article is to describe the problems that arise when one tries to rebuild the documentation of a legacy database and the methods, techniques, and tools through which these problems can be solved. A more in-depth analysis can be found in Hainaut (2002).

BACKGROUND: STATE OF THE ART AND KEY PROBLEMS

Database reverse engineering has been recognized to be a specific problem in the '80s, notably in Casanova and Amaral De Sa (1984), Davis and Arora (1985), and Navathe (1988). These pioneer-

```
Figure 2. A more realistic view of data reverse engineering
```



7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/database-reverse-engineering/20702

Related Content

A Link-Based Ranking Algorithm for Semantic Web Resources: A Class-Oriented Approach Independent of Link Direction

Hyunjung Park, Sangkyu Rhoand Jinsoo Park (2013). *Innovations in Database Design, Web Applications, and Information Systems Management (pp. 1-25).*

www.irma-international.org/chapter/link-based-ranking-algorithm-semantic/74387

A Knowledge-Based Approach to Conceptual Information Retrieval from Full Text

Inien Syuand Sheau-Dong Lang (1991). *Journal of Database Administration (pp. 1-7)*. www.irma-international.org/article/knowledge-based-approach-conceptual-information/51090

Matrix Decomposition-Based Dimensionality Reduction on Graph Data

Hiroto Saigoand Koji Tsuda (2012). *Graph Data Management: Techniques and Applications (pp. 260-284).* www.irma-international.org/chapter/matrix-decomposition-based-dimensionality-reduction/58614

Situational Method Engineering to Support Process-Oriented Information Logistics: Identification of Development Situations

Tobias Bucherand Barbara Dinter (2012). *Journal of Database Management (pp. 31-48)*. www.irma-international.org/article/situational-method-engineering-support-process/62031

Implicit Semantics Based Metadata Extraction and Matching of Scholarly Documents

Congfeng Jiang, Junming Liu, Dongyang Ou, Yumei Wangand Lifeng Yu (2018). *Journal of Database Management (pp. 1-22).*

www.irma-international.org/article/implicit-semantics-based-metadata-extraction-and-matching-of-scholarlydocuments/211912