Efficient Database Evolution in Digital Library Reengineering

Delfina Ramos-Vidal, and Nieves R. Brisaboa

Database Laboratory, Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: delfina.ramos@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.19*

Abstract: With the advancement of internet applications, extensive information systems were created to effectively manage and provide easy access to documents, which coincided with a global initiative to convert physical documents into digital format, making them accessible through the internet. After two decades, these databases are well-structured and organized, although the software used to manage them is gradually becoming outdated. Additionally, once the initial digitization and creation of metadata are completed, it is sensible to enhance the metadata further to provide more detailed information about the documents. In this article we propose a tool to facilitate the evolution of large documentary databases.

1 Introduction

The significant effort to make information public has resulted in huge documentary databases accessible worldwide through digital libraries. Currently, many of these digital library systems or documentary database management systems (documentary DBs) have become technologically obsolete. When considering the reengineering of these systems, apart from updating their software and providing them with new functionalities, there is also a need for the migration of documentary databases. Since the original effort to digitize documents has already been carried out, it is common for the migration process to involve an increase in digitized assets and metadata information to expand the database.

The aim of this article is to propose a tool that facilitates the migration of legacy documentary databases to new data models, automating the process as much as possible in order to save time, both in analysis and implementation.

Over the past twenty years, our research group has developed various digital libraries, such as the Galician Virtual Library[1], the Galician Edition during the Franco era[2], or the Hemeroteca of the Royal Galician Academy. All these platforms require their systems to be updated technologically and to take advantage of this update to provide them with new functionalities. The issue arises from the need to preserve current data while modifying the database schema or integrating new data sources.

Currently, this migration process is based on SQL scripts that import complete tables from the original database, Excel spreadsheets where users can verify and complete data extracted from legacy tables, or CSV files obtained from different external data sources that are intended to be linked with existing data. This work is slow and error-prone since it is not a direct process

[1] Galician Virtual Library: *https://bvg.udc.es/*
[2] Galician Edition during the Franco era: *https://ediciongalizafranquista.udc.gal/*

where the script for data importations must allow domain experts to review original data, which often contain errors after decades of use, fulfil tables with new data they want to introduce into the new database, and make decisions regarding which data to migrate in each case. Based on the experience of migrating these use cases and generalising the real problems and needs observed, we have designed a tool to automate the migration process.

2 State of the art

In the field of databases, issues related to data update, integration, and loading have already been addressed through Extraction, Transformation, and Loading (ETL) tools. These tools were originally designed to move data from operational transactional databases to data warehouses but have been adopted in various other contexts, such as database migration due to system updates. Our initial approach was to investigate whether a standard ETL system could be suitable for our purposes.

Currently, there are ETL solutions available in the market, such as Talend, Pentaho, or Google Cloud Dataflow, among others (Sreemathy et al. (2021)). All of these tools automate most of an organization's workflow without the need for human interaction, providing a highly reliable service, and supporting both on-premise databases and cloud databases. However, the transformations supported by these tools are typically quite generic and limited, including operations like row and column transposition, table joining and splitting, data sorting and filtering, but they do not allow for the definition of domain-specific rules.

Additionally, once a transformation is executed, the user does not have decision-making power to handle errors that may arise, necessitating a subsequent data review step for error correction. In previous work by Ramos Vidal et al. (2022), we presented an initial approach to automating the data transformation process for legacy database migration using Domain-Specific Languages.

3 Our proposal

Figure 1 depicts the architecture of the database migration tool we propose. The primary component of our solution is what we call the "Migration and Collation Tool." Within the tool, three phases can be executed independently. On one hand, the "Model Definition Interface" allows for defining the data model for the data originating from both the Source Database and External Data Sources, as well as the data model for the Destination Database. In this step, the "Migration Rules" specific to the business domain are also defined, which will be checked during the conversion process, and the data model that data requiring expert review should adopt. Additionally, the tool includes a "Data Importer" that enables establishing a connection with both the Source Database and External Data Sources outside the system.

The imported data is processed in the "Converter" component, where data from all sources is evaluated, and necessary transformations are performed based on the migration rules defined earlier. Once converted to the new data model, they are stored in the Destination Database. Data for which the transformation fails or is ambiguous according to the migration rules will be placed in the Review Database. Finally, the records in the Review Database will be displayed in the "Debugging Tool," following the data model for review defined in the previous stage, allowing the user to decide which data to retain in case of ambiguity or what corrective action to take in case of errors before moving them to the Destination Database.

The migration for "Edición en la Galicia franquista" will serve as a running example[3]. The original database (A) contains information about authors with literary production during the Franco era (1936-1975) in Galicia. The goal is to expand the library's corpus with data from an external source (B) covering the period from 1936 to 2000. To automate the migration, we

[3] Edición en la Galicia franquista: *https://ediciongalizafranquista.udc.gal/*

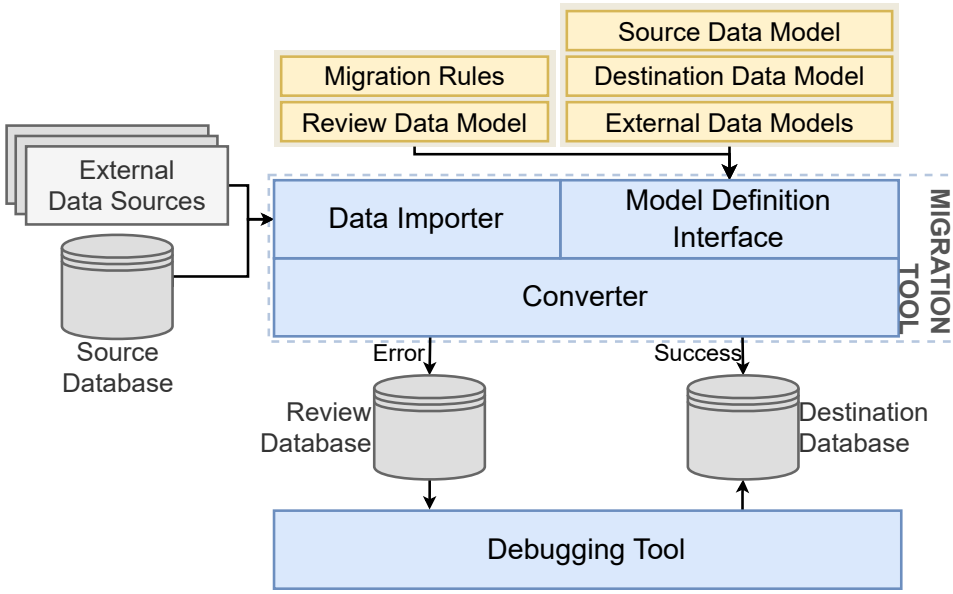

Figure 1: Architecture of the Migration Tool

analyse the possible scenarios that could arise when combining the data sources.

To begin with, if an author is only present in A, the data is automatically migrated to the destination database (C). The same applies when an author has production only after 1975, i.e., they are only present in B. However, in the case of authors with production both before and after 1975, their data will be present in both data sources, so we must evaluate potential cases of inconsistency. We assume that authors in table B are new additions, but a check on table A reveals that there is already another author with the same name, so the row must be sent for review for an expert to verify if they are the same person or not. On the other hand, if the author is present in both tables, we check that all attributes (name, surname, alias, date of birth, date of death, observations) match. It could happen that errors were made when entering authors' names, or an author passed away in recent years, causing a mismatch in the date of death since it doesn't exist in table A. These checks can be defined using rules like:

```
IF (NombreA=NombreB AND AliasA=AliasB)
    THEN (Insert IN C)
ELSE IF (NombreA!=NombreB)
    THEN (Error 1)
ELSE IF (NombreA=NombreB AND AliasA!=AliasB)
    THEN (Error 2)
```

During the conversion process, the attributes defined in the rule are evaluated, and in case of inconsistencies, they are sent to the Review Database for an expert user in the domain to decide which data should be retained or corrected in each case. Once the issues are processed, this data is moved to "C," concluding the migration process.

At the current moment, we are finishing the design phase of the proposal. The next step is to begin the implementation and validation phase, which will include both software-specific verifications and validations, as well as a set of tests with real users in a simulated real-world context.

Acknowledgements

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS), and by PRE2021-099351, MCIN/AEI+"FSE+";GRC[ED431C 2021/53]: GAIN/Xunta de Galicia; TED2021-129245B-C21(PLAGEMIS): MCIN/AEI+"NextGenerationEU"/PRTR; PID2020-114635RB-I00(EXTRACompact): MCIN/AEI; PID2021-122554OB-C33 (OASSIS): MCIN/AEI+EU/ERDF A way of making Europe; PDC2021-120917-C21 (SIGTRANS): MCIN/AEI+"NextGenerationEU"/PRTR

Bibliography

D. Ramos Vidal, A. Cortiñas, M. R. Luaces, O. Pedreira, and A. S. Places. Dsl para la migración de bases de datos legacy en el marco de una lps. In *Actas de las XXVI Jornadas de Ingeniería Del Software y Bases de Datos* (*JISBD 2022*), Santiago de Compostela, 2022.

J. Sreemathy, R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul, and M. Praveennandha. Overview of etl tools and talend-data integration. In *2021 7th International Conference on Advanced Computing and Communication Systems* (*ICACCS*), volume 1, pages 1650–1654, 2021. doi: 10.1109/ICACCS51430.2021.9441984.