Título da comunicação: Database Preservation Toolkit, a relational database conversion and normalization tool

Resumo:

Databases are one of the main technologies supporting organizations' information assets. They are designed to store, organize and explore digital information, thus they are such a fundamental part of information systems that most institutions would not be able to function without them. Very often, the information they contain is irreplaceable or prohibitively expensive to reacquire, thus making the preservation of databases a serious concern.

The Database Management System (DBMS) is the software that manages and controls access to the database, and the database can be described as a collection of related data. These two intrinsically related technologies function together to perform tasks such as information storage and retrieval, data transformation and validation, privilege management and even the enforcement of important business constraints. The most popular databases are based on the relational model[1] proposed by Codd in "*A Relational Model of Data for Large Shared Data Banks*".

The migration of the relational database information into a format well suited for long-term preservation is one of the most accepted strategies to preserve relational databases. This strategy consists in exporting the information of the relational database, including descriptive information, structural information, behavioural information and content, to a format suitable for long-term preservation. Such format should be able to maintain all significant properties of the original database, whilst being a format widely supported by the community and hopefully based on international open standards. Few formats fit this criteria, being the SIARD format one of the main contenders.

[1] According to the ranking at http://db-engines.com/en/ranking (accessed on May 2016), where 7 of the top 10 DBMS use the relational model.

Most DBMSes can export the database to a text file using a format derived from the SQL standard. This format may be proprietary and/or protected by intellectual property rights and thus it is not an ideal database preservation format.

The Software Independent Archiving of Relational Databases (SIARD) format was developed by the Swiss Federal Archives and was especially designed to be used as a format to preserve relational databases. Its second version, SIARD 2, retains the (most commonly agreed upon) databases' significant properties and is based on international open standards, including Unicode (ISO 10646), XML (ISO 19503), SQL:2008 (ISO 9075), URI (RFC 1738), and the ZIP file format.

Since the manual creation of SIARD files is impracticable, an automatic conversion system was developed - the Database Preservation Toolkit. This software can be used to create SIARD files from relational databases in various DBMSes, providing a unified method to convert databases to a database agnostic format that is able to retain the significant properties of the source database. The software uses XML Schema Definition capabilities present in the SIARD format to validate the archived data and can also be used to convert the preserved database back to a DBMS, allowing for some special usage scenarios in an archival context.

However, the digital preservation process stays incomplete if the archived information can not be accessed. To access and explore digitally preserved databases, the E-ARK Database Viewer is being developed. This software can load databases in the SIARD format and display their descriptive, structural and behavioural information and content, enabling a consumer to quickly search and explore a database without knowing any query language. The viewer also provides the functionality to promptly search and filter the database contents as well as export search results, and is able to execute these operations on databases containing millions of records.

The Database Preservation Toolkit and the E-ARK Database Viewer can be used in multiple cases to achieve different goals in an archive context.

During the archiving process, the producer can use DBPTK to convert a database to SIARD 2 and, after adding some documentation and other information, deliver the database to the archive.

After the database is archived, the E-ARK Database Viewer can be used to provide access to the preserved database. A consumer may then use the software to search and filter database records at will.

The E-ARK Database Viewer can also be used to provide controlled access to databases containing sensitive information, as alternative versions of the database can be created that filter sensitive information and/or contain specialized views that omit sensitive information. These prepared versions of the database may then be made available to consumers.

For research purposes, it is also possible to use the DBPTK to convert a preserved database to a system capable of performing data mining or analysis, allowing the full functionality of this system on the preserved dataset.

Nota biográfica:

Bruno Alexandre Alves Ferreira
KEEP Solutions Lda, Braga, Portugal
bferreira@keep.pt

Bruno Ferreira é finalista do Mestrado em Engenharia Informática, na Universidade do Minho, com especializações nas áreas de Processamento de Linguagens e Conhecimento e de Business Intelligence, e realiza a sua tese sobre o tema da preservação de bases de dados. Licenciado em Engenharia Informática pela mesma universidade, Bruno exerce funções como Analista Programador na KEEP SOLUTIONS e está envolvido no projeto E-ARK, um projeto europeu de investigação na área da preservação digital, assim como na conceção e desenvolvimento de ferramentas de preservação digital para a DGLAB. Bruno é atualmente responsável pelo desenvolvimento do Database Preservation Toolkit e do E-ARK Database Viewer.