briefing paper

digital preservation europe

Preservação de Bases de Dados: o desafio internacional e a solução suíça

A maioria dos documentos administrativos é guardada em bases de dados. O actual desafio é preservar a informação e torná-la acessível nos anos futuros, assegurando a transferência de conhecimento bem como a sustentabilidade administrativa. A falta de normalização tem, até agora, prestado a tarefa de arquivar conteúdos altamente complexos. Os Arquivos Federais Suíços desenvolveram um novo formato baseado no XML que permitirá, a preservação a longo termo destes conteúdos relacionais. O Software Independent Archiving Of Relacional Databases (SIARD) que permite uma solução única para preservar estes conteúdos, bem como os metadados, conforme os formatos designados pela ISO.

Porquê arquivo de dados em primeiro lugar?

A preservação de dados a longo prazo sempre foi essencial à administração. Tradicionalmente, assegura planificação e estabilidade. Actualmente a principal assumpção é que os dados informatizados já estão seguros. Arquivar é muitas vezes considerado tão redundante como o nosso hábito de "duplo clique e acesso", que condiciona a maneira como pensamos sobre a preservação digital dos registos.

Contudo, a preservação de bases de dados faz sentido. Primeiro, num ambiente de novas tecnologias sempre em mudança, apenas arquivar poderá verdadeiramente garantir o acesso aos dados e prevenir a sua perda. Segundo, quase 85% dos registos arquivados estão inactivos, representando as actuais bases de dados, muito complexas e dispendiosas de manter. Por último, arquivar é muita vezes prescrito pela Lei, como garantia de liberdade de informação (por exemplo, o Suíço Offentlichkeitsgesetz) ou para documentar actividades governamentais (por exemplo, o francês code du patrimoine). Arquivar garante uma boa resposta às nossas necessidades, preenchendo os requisitos legais, facilitando a gestão de dados e reduzindo os custos operacionais. Contudo, é uma troca difícil e não isenta de armadilhas.

Um contratempo arquivístico, ou o que devemos arquivar?

Uma breve resenha sobre bases de dados ajudar-nos-á a compreender alguns dos principais contratempos no arquivo de base de dados. As primeiras bases de dados (década de 1960) foram organizadas numa hierarquia clara (1:1 ou 1: n relações). Esta estrutura em árvore era propensa a redundâncias, necessárias para que permitissem relações complexas (n:m). Um século depois o modelo hierárquico foi substituído pelo modelo de rede que permite múltiplas relações sem repetições. Mais tarde, outro modelo foi introduzido, o do objecto orientado que consiste num grupo de computadores ligados que representam os dados. Contudo, embora os dados possam ser rapidamente acessíveis o número de consultas neste modelo era reduzido. Para além das diferenças de cada um, os modelos de dados indicados têm um elemento em comum: a dependência dos dados e dos códigos (a linguagem do software dada base de dados). Esta estrita dependência dificulta a extracção de dados da matriz da base de dados. Se não estivermos familiarizados com o código de software dificilmente os poderemos arquivar.

Existe uma excepção à regra: as bases de dados relacionais. Este modelo introduzido por volta de 1970 por Edgar Codd, resolve a dependência de dados e códigos. Armazena todos os dados em tabelas. Esta colecção de tabelas interligadas permite relações múltiplas (n:m) e um numero indefinido de consultas. A utilização de chaves primárias (identificadores únicos para cada fila) e chaves secundárias (referencias às outras tabelas) exime-nos da necessidade de repetições. Ainda que o software se altere, os dados permanecem intactos. Para arquivar, basta-nos extrair e armazenar as tabelas. Arquivar bases de dados relacionais é ainda mais simples, em termos de esforço e custos. Desde que mais de 90% das bases de dados são relacionais, concentrar os nossos esforços na preservação é provavelmente a melhor estratégia. O modelo relacional resolve o nosso maior problema de arquivo. Contudo, é apenas o primeiro passo. O segundo, e talvez o mais delicado é descobrir um formato adequado que assegure o acesos futuro aos dados armazenados. É precisamente isto que os Arquivos Federais Suíços (SFA) estão a tentar fazer.

A solução suíça: o formato SIARD

Os Arquivos Federais Suíços foram confrontados com a questão da preservação de bases de dados desde os finais de 1990. Estrategicamente os SFA decidiram arquivar apenas as bases de dados relacionais. Como parte do projecto ARELDA foi conceptualizado e desenvolvido um novo formato para bases de dados relacionais. O Software – Independent Archiving of Relational Databases (SIARD) foi apresentado em 2004. Desde então tem vindo a ser elaborado e consideravelmente reforçado dentro do projecto PLANETS.

No final do Verão de 2008, o SFA apresentou uma versão completa do formato SIARD com software associado.

Em termos práticos, o que significa preservação a longo prazo com o SIARD? O software SIARD converte as bases de dados proprietárias (MS Access, MS SQL, Oracle, etc. …) num ficheiro de arquivo de formato não proprietário SIARD. O arquivo SIARD (com a extensão de ficheiro.siard) representa a base de dados na sua forma lógica, retendo não apenas o primário e os metadados, mas mais importante todas as relações.

O arquivo SIARD é uma estrutura não comprimida de depósito ZIP (standard ZIP-64), permitindo praticamente ficheiros de todos os tamanhos.

Contém duas pastas: "header" e "content". A pasta "header" armazena o contexto da base de dados, os metadados. Um simples ficheiro, metadata.xml, assegura que poderemos compreender a vertente técnica, bem como o suporte contextual da base de dados. Em termos técnicos SIARD regista ao mais elevado nível (da base de dados), o identificador, a versão formato, a mensagem interpreta o código do pc terminal (verificando a integridade primária dos dados) etc. No nível esquemático SIARD armazena listagens de tabelas, consultas e rotinas. Ao nível das tabelas, o SIARD identifica e regista as falhas. E conforme aprofundamentos o nível SIARD também especifica o SQL em uso, LOBs (Large Objects), nomes, e mais importante de todos: chaves estranhas e candidatos a chave com dados de referência – ou seja, as relações. Ao mesmo tempo SIARD contextualiza dos dados. Ao nível da base de dados permite-nos registar ou adicionar (com o SIARD Suite) informação sobre o arquivo de proveniência, descrição, utilizador, etc.… Aos níveis mais baixos, permite-nos guardar os detalhes dos nomes e conteúdos das tabelas e colunas. Esta informação descritiva torna a base de dados compreensível para utilizadores futuros quer em termos técnicos quer em termos de contexto.

A segunda pasta, "content", armazena os dados primários. Os dados são arquivados de acordo com a estrutura da base de dados. Para cada esquema SIARD automaticamente gera uma pasta (esquema 1, esquema 2, etc…) Os dados em si são armazenados em ficheiros XML (ex: tabela XML 1). A definição deste esquema reflecte os esquemas de metadados das tabelas SQL. E especifica que a tabela é armazenada como uma cadeia de linhas que englobam a sequência de entradas de colunas com diferentes tipos XML. BLOB's e CLOB's (Binary ou Character Large Objects que contém todo o tipo de informação) também são arquivados. São armazenados em pastas geradas automaticamente (por exemplo, lob1, lob 2, etc) seja em ficheiros TXT ou BIN (record1.text, ou record1.bin, etc).

SIARD é o verdadeiro espelho de bases de dados arquivadas. Quando aplicado o SIARD arquivamos ambos, primário e metadados de maneira e na forma que torna a base de dados compreensível e acessível. Mas por quanto tempo?

SIARD e a questão da preservação a longo prazo?

"Eternity "Eternidade é muito tempo.", disse Woody Allen, " sobretudo para o final." No mundo das novas tecnologias uma eternidade poderá ser muito curta. Uma duração de vida efémera ameaça a acessibilidade aos dados em formatos a longo prazo. O que poderá minimizar este risco? Numa palavra: normalização.

A utilização amplamente aceite das normas ISO assegura que o armazenamento de dados seja acessível no futuro. Partindo deste pressuposto os registos dos dados primários e metadados SIARD, tornados automaticamente em formatos de norma ISO: SQL 1999, UNICODE e o mais importante de todos: XML 1.0. Para assegurar a normalização, o SIARD converte todos os dados proprietários para o equivalente conjunto de caracteres UNICODE. Além disso, o SIARD não arquiva sinónimos uma vez que não fazem parte da normalização SQL: 1999. Manter-se fiel às normas é uma regra de ferro.

Por último mas não menos importante

SIARD foi concebido como um formato de ficheiros livre. A sua descrição está disponível no sítio da internet dos Arquivos Federais Suíços. (5) Não pretende ser um canivete suíço para arquivar todos os modelos de bases de dados. Contudo é uma solução viável e prática para a preservação a longo prazo das bases de dados relacionais.

Notes

[1] Yuhanna, Noel "Database Archiving Remains na Important Parto f Enterprise DBMS Strategy", Information & Knowledge Management Professionals (2007): ftp://ftp.software.ibm.com/software/data/sw-library/data-management/optim/reports/forrester-archiving.pdf

[2] ARELDA é o acrónimo para ARchiving of ELectronic Data.

[3] http://www.planets-proejct.eu/

[4] O SFA actualmente utiliza uma aplicação baseada no sistema JAVA, o SIARD Suite, que permite navegar através do arquivo SIARD e adicionar ou actualizar metadados.

Referências

1 - Estratégia de aquisição e disposições dos Arquivos Nacionais (2007) http://www.nationalarchives.gov.uk/documents/acquisitions_strategy.pdf

2 – Codd, E.F., "A Relational Model of Data for Large Shared Data Banks", Communications of the ACM, vol. 13, n.º 6 (1970), 377-387.

3 – Code du Patrimoine, July 30, 2008. http://www.legifrance.gouv.fr/affichCode.do;jsessionid=2FAA76FF7AE923389AC2146821608165.tpdjo01v_2?cidTexte=LEGITEXT000006074236&dateTexte=20081001

4 – Knowles, J.S. / Bell, D.M.R., "The Codasyl Model", in: Databases – Role and structure, P. M. Strocker, P. M. D. Gray, and M. P. Atkinson (eds) CUP, 1984. Swiss Federal Law on Archiving (BGA), June 26 1998: http://www.admin.ch/ch/d/sr/cl152_1.html

5 – Swiss Federal Law on the freedom of information in the federal administration (Offentlichkeitsgesetz, BGO), December 17, 2004: http://www.admin.ch/ch/d/sr/c152_3.html

Author: **Amir Bernstein**, *Swiss Federal Archives*, Amir.Bernstein@bar.admin.ch
Translation: **Sofia Pina,** sofia.opina@gmail.com