ISSN 1981-2841

CRIANDO UM MODELO RELACIONAL PARA O *PUBMED CENTRAL*[1]

CREATING A RELATIONAL MODEL FOR CENTRAL PUBMED

Deivid Patrik Maciel do Santos Fiorin[2], Walkíria Cordenonzi[3] e Giovani Rubert Librelotto[3]

RESUMO

O *PubMed Central* é um repositório de artigos completos publicados, principalmente, nas áreas da saúde, como medicina, biologia, bioinformática e enfermagem. No artigo, descreve-se o processo de construção de um esquema relacional para o *PubMed Central*, tomando como base os esquemas XML, disponibilizados para a criação de uma base de dados local para trabalhos na área de bioinformática.

Palavras-chave: esquemas XML, processo de construção.

ABSTRACT

PubMed Central is a free digital archive of complete published articles in the health field: medicine, biology, bioinformatics and nursing. This paper describes the construction process of a relational modeling to PubMed Central, taking XML Schema as the basis for the creation of a local database to be used in several projects on the bioinformatics field.

Keywords: *XML schema, construction process.*

INTRODUÇÃO

O acesso rápido à informação científica atualizada é de fundamental importância para o profissional da área da saúde. A Biblioteca Nacional de

[1]Trabalho de Iniciação Científica – PROBIC.
[2]Acadêmico do Curso de Sistemas de Informação – UNIFRA. E-mail: deividfiorin@gmail.com
[3]Orientadores – UNIFRA. E-mail: {walkiria,giovani}@unifra.br

Medicina dos Estados Unidos, no final da década de 50, criou o MEDLARS (*Medical Literature Analysus and Retrieval System*), um sistema feito para automatizar a composição da biblioteca médica. Em 1971, a MEDLINE, que significa MEDLARS *On-Line*, começou a oferecer acesso *on-line* a referências da base de dados MEDLARS, mas, apenas em 1997, o MEDLINE foi disponibilizado para todos de forma gratuita.

Um recurso comumente utilizado para pesquisa de literatura científica sobre ciências biológicas é o serviço gratuito, patrocinado pelo NCBI (*National Center or Biotechnology Information* – Centro Nacional de Informações de Biotecnologia) a Biblioteca Nacional de Medicina dos Estados Unidos, conhecido como *PubMed Central*. Esse serviço permite o acesso ao banco de dados MEDLINE, que é um banco de dados de citações referentes à literatura científica sobre ciências biológicas.

De modo a otimizar os algoritmos de busca sobre os dados do *PubMed Central*, este trabalho visa a realização de uma modelagem entidade-relacionamento, a partir do XML Schema (BRADLEY, 2002) dos arquivos disponíveis para *download* do *PubMed Central*. Após a modelagem, os dados dos diversos arquivos XML serão importados para o banco, a fim de se obter um banco relacional local, estruturado da maneira mais adequada.

Conforme os objetivos propostos, o presente artigo está estruturado da seguinte maneira: na seção 2, é introduzido o *PubMed Central*. A criação do modelo relacional, baseado no seu esquema XML, encontra-se na seção 3. Na seção 4, apresenta-se a conclusão deste trabalho, além de se proporem trabalhos futuros.

PUBMED CENTRAL

O *PubMed* é um serviço da Biblioteca Nacional de Medicina dos Estados Unidos (U. S. NLM) que inclui mais de 15 milhões de citações por meio de artigos biomédicos publicados desde 1950, congregando jornais científicos de mais de 40 línguas. O ambiente permite consultas em sete línguas, incluindo inglês, francês e espanhol. A pesquisa pode ser detalhada também por características da amostra do estudo, como idade, gênero, seres humanos, animais, entre outras.

O *PubMed Central* (PMC) é um subconjunto do *PubMed*, consistindo de acesso gratuito para textos na íntegra de mais de 300 mil documentos e cerca de 150 publicações médico-científicas de todo mundo. Portanto, é um repositório digital de acesso livre a jornais e a revistas das áreas biomédicas e ciências da saúde, controlado pela *National Institutes of Health* (NIH, 2007a,b).

O acesso ao *PubMed Central* é feito por meio do endereço http://www.pubmedcentral.nih.gov/. O PMC sempre mostrará o resultado de forma padronizada, ou seja, quase sempre as citações conterão os nomes dos autores, o título do trabalho em inglês, o título da revista e outras informações sobre a publicação, o idioma original do documento, um identificador número do PMC e a situação do documento na base de dados.

ESTRUTURA DOS DADOS NO *PUBMED CENTRAL*

Os artigos, no *PubMed Central,* estão divididos em três partes: cabeçalho, corpo e fim do artigo:

O cabeçalho armazena informações sobre o artigo, o título em inglês (ou traduzido, se já foi publicado em outra língua, mas também guarda o título original em outro campo), informações bibliográficas referentes ao jornal em que o artigo foi publicado (como, por exemplo, o nome da publicação, o título, o ISSN ou ESSN e o volume), dados dos autores, tipo de publicação, a história de uma publicação (por exemplo, recebido, aceito, revisado e publicado) e o *abstract.* Pode conter ainda uma lista de identificadores do artigo, a língua em que foi escrito, o *status* da publicação (impresso ou em meio eletrônico).

O corpo é divido em seções: cada seção possui um título e parágrafos, sendo que, nesses parágrafos, podem haver referências a figuras, tabelas e referenciais bibliográficos contidos no final do artigo. No fim do artigo, encontram-se o referencial bibliográfico, tabelas e figuras, com suas respectivas descrições.

Os dados armazenados no sistema de informação do *PubMed Central* têm uma estrutura rígida e bem definida que pode ser formalizada por uma gramática livre de contexto (MENEZES, 2000), na qual são apresentados apenas os conceitos principais:

Article	==> PMID, Journal, ArticleTitle?, VernacularTitle?, FirstPage, LastPage, Language?, AuthorList?, PublicationType, ArticleIdList, History, Abstract
Journal	==> PublisherName, JournalTitle, Volume, ISSN, Issue?, PubDate
PubDate	==> Year, Month?, Season?, Day?
AuthorList	==> Author+
Author	==> FirstName?, MiddleName?, LastName?, Suffix?, CollectiveName?, Affiliation?
PublicationTypeList	==> PublicationType?
ArticleIdList	==> ArticleId
History	==> received, accepted, revised, aheadofprint

Essa gramática livre de contexto demonstra apenas a estrutura da informação encontrada no *PubMed Central*. Na prática, essas informações estão em banco de dados, estruturados de uma maneira similar.

MAPEAMENTO DOS DOCUMENTOS XML DO PMC PARA O ESQUEMA RELACIONAL

O *PubMed Central* (NLM, 2007b), conforme foi exposto, disponibiliza as informações relativas às suas publicações voltadas aos profissionais da área de ciências biológicas, por meio de arquivos no formato XML (*eXtensible Markup Language*). A quantidade de arquivos utilizados neste trabalho totaliza 51600 arquivos XML, cada um contendo um artigo científico.

Dessa maneira, a pesquisa torna-se uma tarefa árdua, envolvendo toda quantidade de documentos. Um modo de resolver esse problema seria mapear esses dados para um banco de dados relacional, tornando mais viável a pesquisa sobre esses artigos científicos. Sendo assim, o objetivo, neste trabalho, foi realizar o mapeamento dos arquivos XML para um modelo entidade-relacionamento, seguindo, então, para um modelo de banco de dados relacional normalizado.

Para alcançar esse objetivo – modelo relacional – foram executados dois processos em paralelo: um processo denominado de manual e outro utilizando ferramentas; ambos serão descritos a seguir.

PROCESSO AUTOMATIZADO

No processo denominado de automatizado, utilizou-se inicialmente a ferramenta Exult XML (NOVIXYS, 2007). Essa ferramenta faz a leitura da estrutura dos arquivos XML selecionados e cria uma base de dados no banco de dados *Access*.

Com a base de dados definida e os dados inseridos nas tabelas, fez-se a importação pela ferramenta DTS (*Data Transformation Services*) (LARSEN, 2007) para o banco de dados SQLServer (BATTISTI, 2001). Chegou-se a um total expressivo de 182 (cento e oitenta e duas) tabelas. A partir da tabela base, fez-se a engenharia reversa, para obter o modelo entidade-relacionamento, mostrado na figura 1 (apenas uma parte, devido ao número de entidades contidas no modelo).

O problema encontrado em todos esses mapeamentos feitos com ferramentas são as tabelas que não apresentam as restrições de integridade de chave e de integridade referencial, o que torna o modelo inconsistente e praticamente impossível de ser utilizado.

Figura 1 – Parte do Esquema Entidade-relacionamento.

PROCESSO MANUAL

Para se alcançar o modelo entidade-relacionamento – objetivo do trabalho – no processo denominado de manual, um arquivo XML, publicado em cada uma das mais de 100 revistas catalogadas no *PubMed Central,* foi aberto e revisado. Um exemplo dos arquivos XML analisados pode ser visto na figura 2. Para cada *tag* encontrada, foram analisados o seu início e o seu fim, para, então, ser devidamente anotada. Para cada nova *tag,* o processo se repete. Toda vez que já havia a anotação para uma *tag* encontrada, verificava-se a sua cardinalidade, ou seja, se a informação era única ou se repetia. Assim, definiram-se as tabelas, seus atributos e seus relacionamentos.

Figura 2 – Arquivo XML de um artigo do PubMed Central

O resultado obtido é chamado de esquema otimizado, o qual irá gerar o BDOT (Banco de Dados Otimizado). O esquema entidade-relacionamento é mostrado na figura 3. A base de dados foi gerada no *SQLServer* por meio de *scripts sql*.

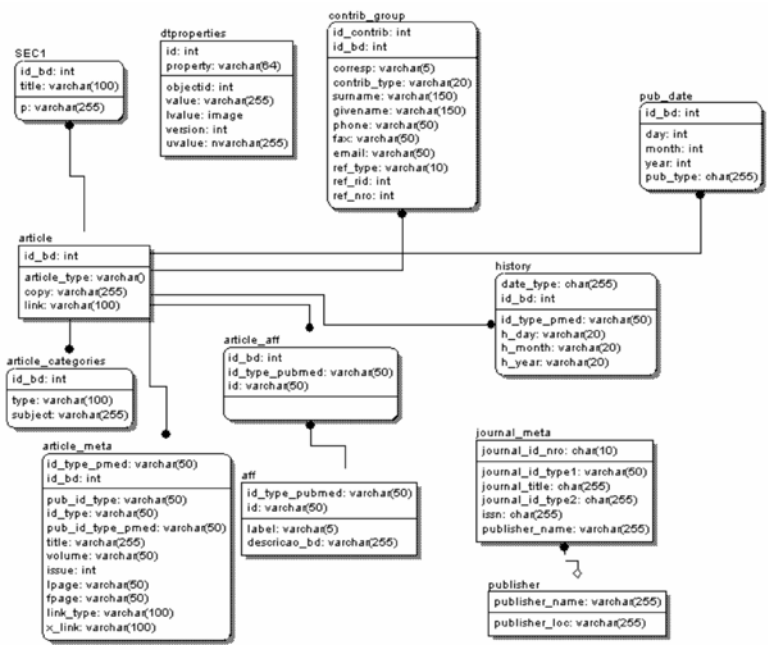

Figura 3 – Base de Dados Otimizada.

A figura 3 mostra o diagrama relacional da base otimizada que foi tratada, além de feitos todos os seus devidos relacionamentos, como os da base original por meio de *scripts* criados no *SQLServer.* A utilização do *SQLServer* ocorre devido ao conjunto de ferramentas existentes no auxílio à construção do modelo.

CONCLUSÃO

A necessidade de dados organizados, de resultados rápidos e corretos e de busca de informação é uma das prioridades dos profissionais que trabalham com bioinformática. Muitas são suas dificuldades e a que este trabalho se propôs a resolver foi a de organizar e agrupar dados para que as consultas às várias bases de dados existentes fiquem mais concisas, organizadas e com um melhor desempenho.

Para atingir o objetivo proposto, fizeram-se duas formas de mapeamento de arquivos XML para o banco de dados *SQLServer,* a fim de que houvesse uma

comparação entre elas. O resultado foi que, ao utilizar ferramentas, ganhou-se tempo para que o banco fosse criado e "populado" com os dados. Em contrapartida, perderam-se as regras de integridade e de referência. Para este banco ser utilizado, é preciso que se faça uma releitura minuciosa dos arquivos para se projetar os *scripts* em que serão definidas as regras. No entanto, o BDOT pode ser considerado como a base para evitar a leitura dos arquivos, pois ela já foi realizada para defini-lo. A desvantagem em utilizar o BDOT é que ele necessita da proteção de um *software* que leia o dado do arquivo XML e coloque na tabela o atributo correto. Além disso, se a estrutura do arquivo for modificada, o *software* não funciona.

O resultado dos dois processos utilizados foi construir *scripts* SQL para transformar a base de dados gerada automaticamente na base otimizada. Os dados continuaram com a integridade da base otimizada, mas com uma complexidade menor, tendo os dados armazenados em um número menor de tabelas, pois, por meio do mapeamento, elas foram reorganizadas de maneira fácil e clara. Com a base de dados final definida e completa (regras e dados), disponibiliza-se uma grande facilidade de manuseio nos dados pelos profissionais da área.

É possível projetar a próxima fase do projeto – definir um *Data Warehouse* (INMON, 1997), com todo esse processo concluído, pois ele tem como finalidade trazer informações históricas possíveis de serem usadas para alguma tomada de decisão ou projeção de alguma meta futura (GALLAS, 2007). Nesse sentido, mostrando como é possível beneficiar todos os usuários que forem usufruir desses dados.

REFERÊNCIAS

BATTISTI, J. **SQL Server 2000 administração e desenvolvimento**: curso completo, Axel Books do Brasil, 2001.

BRADLEY, N. **The XML Companion**, Addison-Wesley, 3 rd edition, 2002.

GALLAS, Susan. **Kimball x Inmon**. Disponível em: <http://www.dwbrasil.com.br/html/dw.html>. Acesso em: jul. 2007.

INMON, W. H. **Como construir o Data Warehouse,** Campus, 1997.

LARSEN, Diane. **Data Transformation Services (DTS).** In: SQL SERVER 2000. Disponível em: http://www.microsoft.com/technet/prodtechnol/sql/2000/deploy/dtssql2k.mspx. Acesso em: jul. 2007.

MENEZES, P. B. **Linguagens Formais e Autômatos**. Porto Alegre: Sagra Luzzatto, 2000.

NIH – U.S. National Institutes of Health. **PubMed.** Disponível em: http://www.ncbi.nlm.nih.gov/sites/entrez. Acesso em: jan. 2007.

NIH – U.S. National Institutes of Health. **PubMed Central –** A free archive of life sciences journals**.** Disponível em: http://www.pubmedcentral.nih.gov/. Acesso em: abr. 2007.

NOVIXYS Software. **Exult XML Conversion Wizard**. Disponível em: http://www.novixys.com/all.html. Acesso em: jun. 2007.