Geração automática de textos em base de dados. Um estudo de caso em ciência e tecnologia

Sandra Regina Martins
Marlon Candido Guérios
José Roberto Tesch Junior
Isabel Maria Barreiros Luclktenberg
Rita de Cássia Romeiro Paulino

Programa de Pós-Graduação em Engenharia de Produção e Sistemas
Universidade Federal de Santa Catarina - UFSC

RESUMO

As interfaces clássicas de sistemas de informações, baseadas em preenchimento de campos, estão fundamentadas em princípios ergonômicos não familiares aos usuários das áreas humanas, que estão mais acostumados com a leitura e interpretação de textos. Neste artigo propõe-se a utilização de algoritmos de geração automática de textos, a partir de conhecimento prévio sobre o conteúdo de bases de sistemas de informações. Mais especificamente, o trabalho apresenta uma metodologia de construção automática de textos a partir de bases de informações em C&T. O resultado é a interpretação do conteúdo da base informado pelo usuário, com impacto positivo em sua própria compreensão das informações fornecidas e principalmente na qualidade dessas informações.

Palavras-chave: Sistemas de informações, base de dados relacional, áreas humanas, geração automática de textos, qualidade da informação, linguagem natural.

ABSTRACT

Principles of Human Computer Interface applied on the development of many information systems are mostly based on ergonomic concepts unfamiliar for users coming from the linguistic area. These users are far more adapted to read and comprehend complete texts than to understand the semantics of different fields presented on a graphical user interface. This article explores the use of text generation algorithms in order to translate information presented in relational databases to a natural phrasal structure. It is also introduced a methodology to apply these translation approach over databases containing scientific and technical information. The expected result of the proposed work is to facilitate the comprehension of scientific and technical information to users not closely related to these areas.

Keywords: Information systems, relational database, text Generation algorithms, human computer interfaces, natural language.

1. INTRODUÇÃO

Com a introdução de novas tecnologias de informação em diversos campos do conhecimento, sobretudo no campo das ciências humanas, cada vez mais se evidencia uma constante preocupação com a transferência da informação oriunda dessas tecnologias. Nas ciências humanas, a interação comunicativa por meio de textos é mais comum do que, por exemplo, a que se estabelece via interfaces de Sistemas de Informação (SI). Ou seja, como essas interfaces apresentam informação de forma estruturada, baseando-se em preenchimento de campos, exigem dos seus usuários um maior esforço cognitivo em atividades de interpretação e expressão das informações que o sistema processa.

A internet é um canal barato e confiável de distribuição, que permite que as empresas tornem as informações disponíveis para todos os usuários(empregados, clientes, fornecedores). Os canais de distribuição incluem internet, intranet e *broadcasting*. Um elemento crítico para o sucesso dos portais nas corporações é a acessibilidade de informações por todos os usuários. As empresas precisam usar técnicas para garantir que a informação certa está disponível e distribuída para a pessoa certa, no momento certo. As aplicações de portais irão aumentar as oportunidades de mercado para as áreas de Gerenciamento de Conteúdo, Fornecedores de Banco de Dados e *Business Intelligence*. [Lynch,1998]

No processo resultante dessa interação usuário–sistema há uma preocupação manifesta por parte do usuário com relação à qualidade e precisão das informações fornecidas, principalmente porque em interfaces de formulário o sistema requer diferentes categorias de informação, e alguns tipos de dados devem ser digitados muitas vezes, o que aumenta a possibilidade de ocorrerem inconsistências.

Considerando-se que a 'leitura' de informações de natureza estruturada e relacional pode ocasionar problemas à fidedignidade das informações fornecidas pelo usuário de SI, propõe-se que os dados capturados por esses sistemas sejam apresentados em forma de texto descritivo, o que pode contribuir para melhorar a qualidade desses dados, tornando-os mais fidedignos ao que foi informado.

Assim sendo, este artigo tem por objetivo apresentar uma metodologia de geração automática de textos a partir de informações presentes em base relacional, que, por sua vez, pode ser descrita como a base de qualquer sistema computacional.

O domínio no qual estará concentrado este trabalho refere-se ao contexto de informações curriculares. A aplicação da metodologia em questão consiste nos sistemas de captura de uma Plataforma Nacional de Sistemas de Informações em Gestão de C&T, a Plataforma Lattes. Essa plataforma foi concebida para ser uma ferramenta que provê informações a usuários, agências de fomento e instituições de ensino e pesquisa. Os sistemas que compõem a Plataforma Lattes atuam sobre uma ampla base de dados aberta e compartilhada, permitindo assim que seus usuários atuem como beneficiários e provedores de informação. E como a qualidade de um sistema de informação depende do projeto do Banco de Dados (BD), é necessário considerar a modelagem da informação, que é essencial na etapa do projeto de BD. Esse item será descrito na próxima seção.

2. SISTEMAS DE INFORMAÇÕES E MODELAGEM DA INFORMAÇÃO

A modelagem da informação trata da concepção da estrutura da informação em determinado universo do discurso (ambiente, empresa, negócio, etc.) (KERN; RAMOS, 2002). Para o que se propõe neste

artigo, a modelagem é fundamental, visto que implica diretamente no projeto do BD e, por conseqüência, no teor da informação que ali se encontra. Estabelecendo-se uma correlação com o resultado a que se pretende chegar com a extração de informações da base de dados, devem ser considerados os seguintes pontos:

- há semelhanças entre linguagens de modelagem de BD e linguagem natural (LN);
- existe um texto embutido na **estrutura** e no **conteúdo** de um BD;
- pode-se formar frases a partir da estrutura do BD (metadados) e do conteúdo (dados).

Na literatura consultada sobre linguagem natural, referenciada na bibliografia, são registradas semelhanças entre as sintaxes de linguagens naturais e linguagens de modelagem de BD. Entidades são substantivos (que funcionam como sujeito ou objeto direto), relacionamentos são verbos transitivos, e atributos são adjetivos.

A semelhança, no entanto, muitas vezes passa despercebida por modeladores e peritos no domínio do negócio representado no BD. As linguagens naturais são flexíveis e permitem sentenças ambíguas, enquanto as linguagens de modelagem são rígidas e não-ambíguas, semelhantes à lógica de primeira ordem.

É possível, em um BD, construir texto tanto a partir dos metadados (dados sobre os dados, esquema do BD) quanto a partir dos dados. Quanto aos metadados, são fontes de texto: o glossário de entidades, as frases que expressam relacionamentos e a lista de atributos ou características de cada entidade.

Em suma, o próprio desenho de um esquema de BD pode ser encarado como uma tarefa de redação. A população de um banco com dados cria novas sentenças, particularizando as sentenças relativas aos metadados. E como o contexto do artigo em questão envolve aspectos textuais, na seção seguinte serão enfocadas algumas considerações acerca do termo “texto” bem como de pontos referentes à geração automática de textos verificados em alguns papers, referenciados na bibliografia deste trabalho.

3. GERAÇÃO AUTOMÁTICA DE TEXTOS

Para se abordar a questão sobre a geração automática de textos, cabe enfatizar algumas definições conceituais acerca do termo ‘texto’.

Conceituando “texto”. De maneira tradicional, entende-se por texto um conjunto de enunciados que se inter-relacionam formando um todo significativo. O texto não se caracteriza como um aglomerado de palavras ou frases dispostas aleatoriamente (FIORIN e SAVIOLI, 1991). Bem mais que isso, constitui um “tecido” em que uma informação se atrela à outra para formar o sentido.

Quando se fala em texto, não se pode deixar de mencionar a questão da leitura e compreensão. Para entender um texto, o leitor deve considerar os três significados distintos do verbo “compreender” – conter em si, entender/interpretar e aprovar.

Quanto ao primeiro sentido, o leitor deve entender o conjunto de caracteres que definem um conceito, ou seja, que determinam o significado do texto propriamente dito. O segundo significado desse verbo está relacionado ao conceito de interpretação/apreensão do sentido do texto, interpretação neste caso ligada à descrição do sistema de signos de maneira formal, não permitindo ambigüidades

e/ou interpretações múltiplas. E o último sentido une-se à idéia de aceitação pelo leitor das informações contidas no texto.

Em suma, para compreender o texto, o leitor deve entender o seu significado, interpretá-lo e aceitá-lo.

Relação entre coesão e coerência. Partindo-se do pressuposto, segundo alguns estudos, de que a coerência é fator fundamental da textualidade, ela constitui um princípio de interpretabilidade, que é o processo cooperativo entre produtor e interlocutor para que possa haver compreensão (KOCH e TRAVAGLIA, 1990). É o que faz com que o texto faça sentido para os seus leitores. A coerência depende de uma intrincada rede de fatores de ordem lingüística, semântica, cognitiva, pragmática e interacional. Para haver coerência, é preciso que haja possibilidade de se estabelecer no texto alguma forma de unidade ou relação entre seus elementos.

Paralelamente ao conceito de coerência, encontra-se nos estudos textuais a coesão, definida como a organização articulada entre os vários enunciados do texto, a concatenação entre eles (FIORIN e SAVIOLI, 1991). A coesão permite que os elementos lingüísticos presentes no texto se interliguem e formem seqüências veiculadoras de sentido (KOCH e TRAVAGLIA, 1990). O conectivo é um elemento considerado fundamental para que haja coesão no texto. É o que vai permitir uma ligação lógica entre termos e orações de um período.

A relação entre coesão e coerência existe porque a coerência é estabelecida a partir da seqüência lingüística que constitui o texto, isto é, os elementos dispostos na superfície textual servem de pista para que a coerência se concretize.

Geração automática de textos e sumarização. Nesse contexto, também é importante diferenciar o processo de geração automática de textos de outro mecanismo semelhante, mas com objetivo bastante distinto: a sumarização (RINO, 1996). Este outro procedimento visa extrair de um texto acabado suas informações principais. Como exemplos típicos para ilustrar o emprego de sumarização, pode-se citar os “clippings” de notícias exibidos em portais na internet, além de informações sumarizadas enviadas para telefones móveis (SMS, WAP). Já, para ilustrar as aplicações de geração automática de textos, pode-se citar a documentação automática de sistemas, a elaboração automática de cartas, a geração de relatórios e até a criação de subsistemas de ajuda ao usuário.

Fatores envolvidos na geração de textos. Sabe-se que a escrita de textos envolve processos cognitivos complexos. E como não é tarefa fácil para boa parte das pessoas, resulta de aprendizado constante, busca por aperfeiçoamento, esforço e aplicação. E, em se tratando de geração automática de conteúdo textual, há que se avaliarem fatores diversos para se atingir esse fim.

O processo de geração automática de texto é dividido normalmente em duas partes distintas (BOYER; LAPALME, 1990): (1) "o que dizer”, que extrai os fatos importantes da informação disponível; e (2) "como dizer", que permite definir técnicas de retórica apropriadas ou escolhas sintáticas para expressar a informação previamente selecionada. Para os autores, a segunda parte representa uma limitação, já que, por exemplo, em poesia, é possível ter mais interesse pelo efeito no ouvinte ou em alguns aspectos do canal de comunicação do que pelo conteúdo de mensagem propriamente dito. Entretanto, como na

maioria das aplicações práticas é a função referencial do texto que prevalece, a maior preocupação no que diz respeito à geração automática de textos concentra-se em "o que dizer".

Remetendo-se ao apontamento feito anteriormente na seção 2 sobre linguagem natural e sua semelhança com modelagem de BD, cabe descrever a LN como qualquer linguagem de uso geral, escrita e/ou falada por uma comunidade humana. Ao passo que um processo de sumarização permite a criação de abstracts ou sumários a partir de um texto na forma como foi redigido pela fonte humana, a geração automática de textos permite a criação de textos a partir de uma representação computacional (WOOLLEY, 19–). A geração automática de textos pode então ser definida como a geração de textos de linguagem natural, através da utilização de ferramentas computacionais.

A abordagem aqui tratada refere-se à capacidade de se gerar conteúdo textual a partir de um BD. Uma estrutura de BD (grosso modo, semelhante a uma estrutura de árvore), através da qual o texto seria elaborado, tornaria isso possível de uma maneira mais precisa, pois as relações entre os elementos do BD correspondem a orações e relacionamentos entre as orações e os grupos de orações.

4. MÉTODO PROPOSTO E APLICAÇÕES

Pré-requisitos do método. O método para a geração de textos aqui proposto baseia-se na criação de um template, isto é, um documento que contém informações personalizadas que descrevem os dados, as quais serão utilizadas para a geração do texto e que define a sua ordem e como este será apresentado. Isso significa que é necessário que se tenham metainformações sobre os dados da aplicação-fim. A construção do catálogo dessas metainformações deve levar em consideração a existência de um especialista na área-fim do sistema (base de currículos), de modo que essa pessoa sugira um glossário de possíveis palavras que serão utilizadas para gerar o texto, para que ele seja coerente e apresente consistência em relação ao seu contexto de uso. Assim se deve supor que o especialista possua o mínimo de conhecimento em relação à modelagem de dados do sistema.

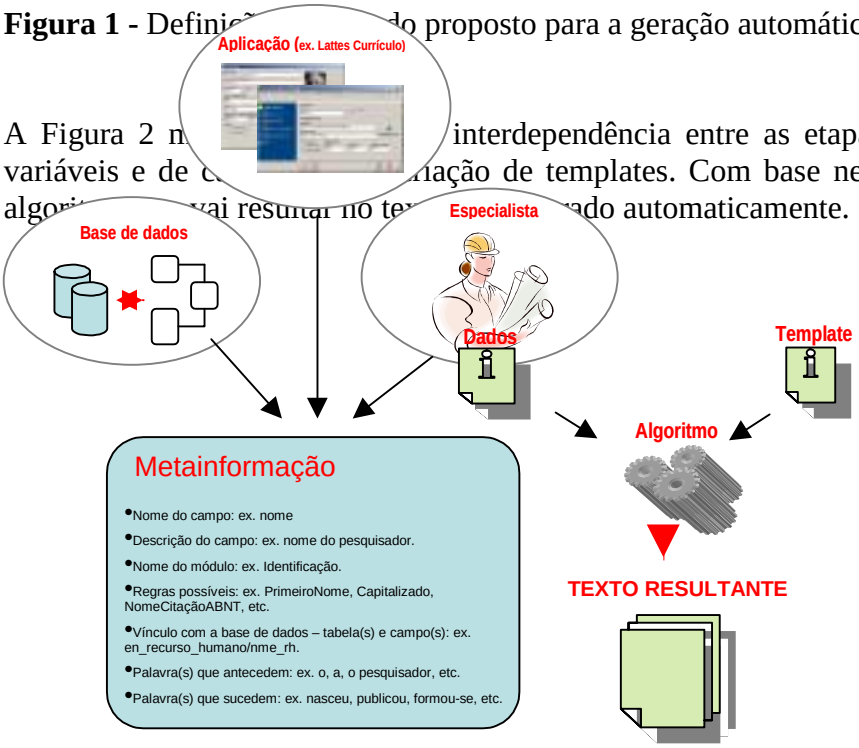
Definição do método. O método consiste nas seguintes etapas, descritas na seqüência:

- (1) cadastro e manutenção de variáveis e de catálogo: nesse contexto, as *variáveis* seriam metainformações que relacionam os campos das interfaces da aplicação-fim com a base de dados. Tais variáveis representam um padrão de texto possível de ser utilizado em um texto da área, no contexto em questão, o conteúdo voltado para informações curriculares. Possuem uma identificação clara e informações que dizem em qual contexto devem ser utilizadas. Em síntese, as variáveis correspondem a um conjunto de palavras que podem anteceder e um conjunto de palavras que podem suceder ao texto fornecido pela variável. Já o catálogo propriamente dito seria o conjunto de todas as variáveis catalogadas do sistema de informações, ou seja, metainformações que receberiam a devida classificação com o auxílio de um especialista da área.
- (2) criação de templates: nesta etapa, uma lista de variáveis disponíveis no catálogo seria ajustada para o tipo de texto da área-fim do sistema.
- (3) aplicação do algoritmo: esta etapa corresponde à aplicação do algoritmo que, a partir de cada variável encontrada no template, relaciona as metainformações nelas contidas com a base de dados da aplicação-fim.

A Figura 1, apresentada a seguir, mostra as etapas necessárias à definição do método proposto para gerar textos automaticamente. Ou seja, na criação do catálogo em que irão constar as variáveis para a

geração do algoritmo do qual vai resultar o texto final, é necessário considerar (a) a aplicação-fim – no caso em questão o Sistema de Currículos Lattes –, (b) o modelo do BD dessa aplicação e (c) um especialista da área que possua conhecimentos em modelagem.

Figura 1 - Definiç do proposto para a geração automática de textos



A Figura 2 n interdependência entre as etapas de (a) cadastro e manutenção de variáveis e de c nação de templates. Com base nessas duas etapas iniciais, é criado o algori vai resultar no te ado automaticamente.

Figura 2 - Algoritmo para a geração de texto baseado em template

Esse resultado textual final pode ter utilidades diversas, como os exemplos de aplicação que se seguem.

(a) Resumé de usuário, com base no preenchimento do CV-Lattes

A partir da especificação do sistema, do modelo de dados e de regras aplicáveis, além dos elementos de sintaxe disponíveis, é possível construir regras sintáticas de formação de frases que, quando combinadas, geram parágrafos explicativos sobre as informações que constam nas bases do sistema descrito. O exemplo a seguir inclui a descrição em resumé do perfil profissional de um determinado pesquisador:

João da Silva nasceu em 1960 na cidade de Dusseldorf, Alemanha. Ele se graduou em Psicologia na Universidade de Wuppertal, Alemanha, em 1983 […]

Avaliação das sentenças

1ª Frase:
João da Silva nasceu em 1960 na cidade de Dusseldorf, Alemanha.
<CAMPO>+<VERBO>+<PREPOSIÇÃO>+<CAMPO>+ <PREPOSIÇÃO>
Módulo: IDENTIFICAÇÃO Módulo/Campo: Identificação: cidade nascimento + Módulo: IDENTIFICAÇÃO
Campo: Nome: **JOÃO DA SILVA** + **NASCEU** + **EM** + Campo: Data Nascimento: **1960** + **EM**
Regra: Valor completo + Regra: Extrair Ano de DD/MM/AA

<CAMPO> + <PONTUAÇÃO> + <CAMPO> + <PONTUAÇÃO>
Módulo: IDENTIFICAÇÃO + Módulo: IDENTIFICAÇÃO
Campo: Cidade Nascimento: **DUSSELDORF** + ‘,’ + Campo: País de Nascimento: **ALEMANHA** + ‘ **.** ’
Regra: Valor completo + Regra: Valor Completo

2ª Frase:
Ele se graduou em Psicologia na Universidade de Wuppertal, Alemanha, em 1983 […]
<PRONOME>: (selecionar da base + Regra do Módulo/Campo = ‘Identificação/Sexo’): **ELE**
+ <VERBO>: (selecionar Módulo/Campo): Formação/Nível + Regra = ‘graduação’: **GRADUOU**
+ <PREPOSIÇÃO>: (selecionar da base): **EM**
+ <CAMPO>: Módulo/Campo (Formação/Área do Curso): **PSICOLOGIA**
+ <PREPOSIÇÃO>: (selecionar da base): **NA**
+ <CAMPO>: Módulo/Campo (Formação/Instituição do Curso): **UNIVERSIDADE DE WUPPERTAL**
+ <PONTUAÇÃO>: (selecionar da base): **,**
+ <CAMPO>: Módulo/Campo (Formação/País da Instituição do Curso): **ALEMANHA**
+ <PONTUAÇÃO>: (selecionar da base): **,**
+ <PREPOSIÇÃO>: (selecionar da base): **EM**
+ <CAMPO>: Módulo/Campo (Formação/Ano de Conclusão do Curso): **1983**

+ <PONTUAÇÃO>: (selecionar da base):

(b) Declarações (statements) sobre o perfil profissional do usuário, a partir das informações indicadas em seu currículo

A riqueza dos detalhes atualmente presentes no Sistema CV-Lattes permite a construção de um ferramental inédito em termos de descrição textual automática de currículos. Os exemplos "O Sr. Usuário Fulano de Tal publicou seu primeiro trabalho internacional aos 41 anos", "A Sra. Beltrana de Tal tem 20 anos de experiência em cirurgia cardiovascular" e "Nos últimos 15 anos, o Sr. Cicrano de Tal orientou oito mestrandos em Engenharia Elétrica, um em Ciência da Computação e quatro doutores na área de Engenharia Elétrica" são perfeitamente dedutíveis de processo "template", a partir das bases de dados Lattes. Deve-se notar que a discordância por parte do usuário poderá levá-lo a alterar informações equivocadas ou a incluir dados faltantes (ex.: inclusão do primeiro artigo publicado para corrigir o tempo sobre quando publicou o primeiro trabalho internacional.). Acredita-se que a médio prazo haverá contribuições significativas para a qualidade dos dados disponíveis na base.

(c) Regras de controle e auditoria de qualidade de informação, aplicáveis ao Sistema CV-Lattes

Considerando-se que os usuários do Sistema CV-Lattes se preocupam com a qualidade e precisão das suas informações curriculares, pretende-se gerar uma ferramenta que seja integrada ao Sistema de Currículos Lattes e que permita aos autores de currículos Lattes obterem descrição textual de sua vida curricular, a partir de métodos de extração e escrita das informações constantes nas bases do Sistema CV-Lattes.
Abaixo constam algumas regras que poderiam verificar a qualidade dessas informações:

I. MÓDULO IDENTIFICAÇÃO

1. O Sr. João Santos da Silva utiliza a abreviatura SILVA, J. S. em suas produções científicas e acadêmicas
   a. Ações do Sistema:
      i. 'Sr': indicar a possibilidade de erro no campo 'Sexo' (com envio para o campo, se o usuário desejar)
      ii. 'Nome': indicar a possibilidade de erro no campo 'Nome' (idem)
      iii. 'Abreviatura': idem

2. O Sr. João Santos da Silva é capixaba, natural de Vila Velha
   a. Ações do Sistema:
      i. 'Sr': indicar a possibilidade de erro no campo 'Sexo' (com envio para o campo, se o usuário desejar)
      ii. 'capixaba': a partir de um glossário de naturalidade (ES = 'capixaba'), o sistema pode indicar o Estado de nascimento, e, caso discorde, o usuário pode trocar (como em alteração de erros)
      iii. 'Vila Velha': campo Cidade de Nascimento (idem)

II. IDIOMAS (E PRODUÇÃO C&T)

1. Dos 3 idiomas que o Sr. João Santos da Silva registra proficiência, há produção científica e tecnológica em dois idiomas registrada em seu CV

a. Incluir Idiomas
b. Alterar Idiomas nas produções
c. Incluir Produção

2. O Sr. João Santos da Silva publicou seu primeiro artigo em revista aos 30 anos de idade
   a. Alterar Idade
   b. Incluir Primeiro Artigo em Revista

5. TECNOLOGIA XML

Introdução. Como mencionado anteriormente, este projeto se baseia em uma base de dados, que neste caso são arquivos XML onde se encontram os dados e metadados definidos na LMPL (Linguagem de Marcação da Plataforma Lattes). Além disso, o template necessário para a geração automática de texto será fundamentado na tecnologia XSL. Do processamento do arquivo XML de um currículo com um arquivo de template XSL resultará uma página com as informações presentes no currículo de maneira descritiva, em uma linguagem discursiva tradicional que poderá ser mais facilmente compreendida pelo usuário.

XML. Conforme afirma o W3C (World Wide Web Consortium), o padrão XML (Linguagem de Marcação Extensível) descreve uma classe de dados chamada documentos XML e parcialmente o comportamento de programas de computador que os processa. O XML é um perfil de aplicativos ou forma restrita do SGML (Linguagem de Marcação Generalizada Padrão): “Em termos construtivos, os documentos XML são documentos que atendem ao padrão SGML” (W3C, 2001).

Segundo o W3C, os documentos XML são formados por unidades de armazenamento chamadas entidades, que contêm dados analisados sintaticamente ou não. Os dados analisados sintaticamente são constituídos por caracteres, alguns dos quais formam dados de caracteres e parte dos quais formam a marcação. Esta última codifica uma descrição do armazenamento e da estrutura lógica do documento. O XML fornece um mecanismo para impor restrições no layout de armazenamento e da estrutura lógica.

Um módulo de software chamado processador XML é usado para ler documentos XML e fornece acesso a seus conteúdos e estruturas. Assume-se que um processador XML esteja fazendo seu trabalho em nome de outro módulo, chamado aplicativo. Essa especificação descreve o comportamento exigido de um processador XML ao ler os dados XML e fornecer as informações ao aplicativo.

O XML foi desenvolvido por um grupo de trabalho XML (originalmente conhecido como Comitê de Revisão Editorial do SGML), formado sob os cuidados do W3C, em 1996. Ele foi presidido por Jon Bosac da Sun Microsystems, com a participação ativa de um grupo de interesses especiais para XML (previamente conhecido como grupo de trabalho SGML), também organizado pelo W3C. No caso da Plataforma Lattes, optou-se por gerar um arquivo XML para cada currículo com o intuito de possibilitar e facilitar a integração com outros sistemas.

XSL. Segundo Elizabeth Castro (2001), o XML em si é simples, o que o torna poderoso são as tecnologias que são utilizadas juntamente com a linguagem. Dispondo-se de um documento estruturado em XML, é possível utilizar ferramentas específicas, conhecidas como padrões companheiros, para

manipulá-lo da forma desejada. Essas tecnologias também são padrões criados pelo W3C para serem utilizados em documentos XML. Cada um possui sua especificação na página do W3C.

O XSL é o mecanismo de folhas de estilo personalizado para o XML. O objetivo das folhas de estilos é informar aos programas que irão interpretar os documentos como exibi-los. Segundo Castro (2001), a especificação do XSL deveria conter a proposta completa e oficial para transformar e formatar documentos XML. Porém, como essa proposta estava demorando muito para ficar pronta, o W3C resolveu dividir o XSL em duas partes: XSLT (XSL Transformation) para transformação de documentos XML e XSL-FO (Formating Objects) para formatação de objetos.

O XSL-FO ainda não possui sua especificação formalizada pelo W3C e nem é compatível com nenhum tipo de navegador da internet. Já o XSLT teve sua especificação publicada no site do W3C em 16 de novembro de 1999.
Pode-se usar o XSLT para transformar os documentos XML em documentos RTF (Rich Text Format), outros documentos XML ou, o que tem sido mais comum, converter documentos XML em HTML para serem visualizados nos navegadores da internet. De acordo com Castro (2001, p. 135), "transformar um documento XML significa analisar seu conteúdo e tomar determinadas ações dependendo dos elementos encontrados. Pode-se utilizar a linguagem para reordenar a saída de acordo com determinado critério, para só exibir determinadas partes da informação, e muito mais".

Aplicação de XML e XSL na Geração Automática de Textos. Como se pode observar na Figura 3, através de parte de um arquivo em formato XML é possível perceber que essa não é a melhor forma de se visualizarem as informações curriculares de um indivíduo.

```
- <DADOS-GERAIS NOME-COMPLETO="Fernando da Silva" NOME-EM-CITACOES-BIBLIOGRAFICAS="SILVA, F." NACIONALIDADE="B" CPF="82254444772" PAIS-DE-
  NASCIMENTO="Brasil" UF-NASCIMENTO="RJ" CIDADE-NASCIMENTO="Rio de Janeiro" DATA-NASCIMENTO="07091963" SEXO="MASCULINO" NUMERO-IDENTIDADE=""
  ORGAO-EMISSOR="" UF-ORGAO-EMISSOR="" DATA-DE-EMISSAO="" NUMERO-DO-PASSAPORTE="" NOME-DO-PAI="        " NOME-DA-MAE="              " PERMISSAO-DE-
  DIVULGACAO="SIM" OUTRAS-INFORMACOES-RELEVANTES="">
```

Figura 3 - Parte de um currículo Lattes em formato XML

Devido a isso, é proposta a utilização de um template XSL que irá transformar esse arquivo de maneira que as informações fiquem facilmente compreensíveis pelo usuário que não conhece a definição dos metadados utilizados na criação do arquivo XML.

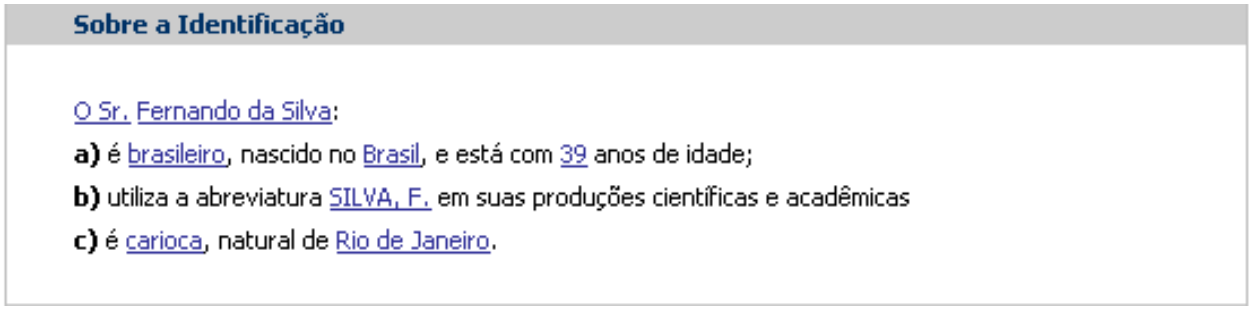


Figura 4 - Apresentação de parte de um arquivo XML transformado com XSL

São perceptíveis na Figura 4 os benefícios alcançados com a transformação do arquivo XML com o template proposto.

6. CONCLUSÕES

Como resultado dos avanços tecnológicos ocorridos nos últimos tempos, é possível ter acesso de maneira facilitada a uma quantidade muito grande de informações. Porém, nem sempre uma informação "acessível" significa uma informação com a qualidade ou confiabilidade desejadas. Mesmo com a grande diversidade de sistemas de informação existentes, voltados normalmente para o âmbito empresarial (e não acadêmico), percebe-se que tais sistemas pouco têm evoluído em sua dimensão humana, isto é, aquela que permite interação direta com as pessoas. A importância desse enfoque deve-se, sobretudo, à característica do usuário do sistema, o qual pode ser desde um executivo, com pouco tempo disponível para procurar ou mesmo filtrar a informação que realmente é de seu interesse, até um usuário acadêmico, com pouco conhecimento de sistemas de informação e formulários eletrônicos, que apenas deseja obter a informação em sua forma textual tradicional, com a qual está habituado. Cada processamento adicional que o usuário necessite acrescentar à informação que lhe é apresentada implica em comprometer a clareza ou precisão dessa informação, uma vez que qualquer interpretação está sujeita ao contexto e nível de conhecimento do usuário sobre o que lhe é apresentado. Nesse panorama, o processo de geração automática de textos possui um papel fundamental, pois a extração de informações de uma base de dados, mesmo sendo muito bem estruturada, e sua posterior disposição na forma escrita tradicional envolvem fatores e processos bastante complexos. Uma coisa é certa: o processamento adequado da informação sempre foi e continuará a ser um grande desafio para qualquer tipo de sistema de informações.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALBRECHT, M. E. et al. An Informal and Efficient Approach for Obtaining Semantic Constraints using Sample Data and Natural Language Processing, In: PROCEEDINGS OF THE INTERNATIONAL WORKSHOP SEMANTICS IN DATABASES, Prague, 1995.

BOYER, M.; LAPALME, G. Text Generation. In P. Saint-Dizier and S. Szpakowicz, editors, Logic and Logic **Grammars for Language Processing**, chapter 12, pages 255-274. Ellis Horwood, 1990.

CHEN, P. P. English sentence structure and Entity-Relationship diagram. **Information Science** 29 (2), Elsevier, pp. 127-149, May 1983.

FIORIN, J. L; SAVIOLI, F. P. **Para entender o texto**. São Paulo: Ática, 1991.

KERN, V. M.; RAMOS, A. L. M. Bridging the gap between natural and information modeling languages: an informal approach to information modeling learning, In: Proc. of the 7th Intl. Conf. on Engineering and Technology Education – Intertech'2002 (CD-ROM), 5 p., Santos-SP, Brazil, March 17-20, 2002.

KOCH, I. G. V; TRAVAGLIA, L. C. **Texto e coerência**. São Paulo: Contexto, 1990.

National Institute of Standards and Technology (NIST), Federal Information Processing Standards Publication 184, Integration Definition for Information Modeling (IDEF1X), Gaithersburg, MD, December 1993.

PRISCILA, Walmsley. **Definitive XML Schema**. Upper Saddle River, NJ: Prentice-Hall, Inc. 2002.

RINO, L. H. M. A sumarização automática de textos em português. In: ACTAS DO II ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DE PORTUGUÊS ESCRITO E FALADO (PROPOR'1996*)* (Curitiba, Paraná, Brasil, 21 e 22 de outubro de 1996), p.109-119.

TESCH JR., José Roberto. **XML Schema**. Florianópolis: Visual Books, 2002.

WOOLLEY, G. H. **Automatic Text Generation**. Manager of Sigma 5/7, Time Sharing Development, Scientific Data Systems. (A Xerox Company), 701 South Aviation Boulevard. EI Segundo, California 90245 - U.S.A., 19–.