

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**UM PROCESSO AUTOMATIZADO PARA
TRATAMENTO DE DADOS E CONCEITUALIZAÇÃO
DE ONTOLOGIAS COM APOIO DE VISUALIZAÇÃO**

ELIS CRISTINA MONTORO HERNANDES

São Carlos - SP
Agosto/2009

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**UM PROCESSO AUTOMATIZADO PARA
TRATAMENTO DE DADOS E CONCEITUALIZAÇÃO
DE ONTOLOGIAS COM APOIO DE VISUALIZAÇÃO**

ELIS CRISTINA MONTORO HERNANDES

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software.
Orientadora: Dra. Sandra Carmargo Pinto Ferraz Fabbri.

São Carlos - SP
Agosto/2009

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

H557pa

Hernandes, Elis Cristina Montoro.

Um processo automatizado para tratamento de dados e
conceitualização de ontologias com apoio de visualização /
Elis Cristina Montoro Hernandez. -- São Carlos : UFSCar,
2010.

169 f.

Dissertação (Mestrado) -- Universidade Federal de São
Carlos, 2009.

1. Engenharia de software. 2. Metadados. 3. Avaliação
educacional. 4. Instituto Nacional de Estudos e Pesquisas
Educaçãois (INEP). 5. Ontologia. 6. Visualização. I. Título.

CDD: 005.1 (20^a)

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

**“Um Processo Automatizado para Tratamento
de Dados e Conceitualização de Ontologias com
Apoio de Visualização”**

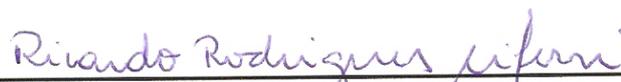
ELIS CRISTINA MONTORO HERNANDES

Dissertação de Mestrado apresentada ao
Programa de Pós-Graduação em Ciência da
Computação da Universidade Federal de São
Carlos, como parte dos requisitos para a
obtenção do título de Mestre em Ciência da
Computação

Membros da Banca:



Profa. Dra. Sandra C. Pinto Ferraz Fabbri
(Orientadora - DC/UFSCar)



Prof. Dr. Ricardo Rodrigues Ciferri
(DC/UFSCar)



Prof. Dr. Ricardo de Almeida Falbo
(INF/UFES)

São Carlos
Agosto/2009

*Para realizar certos sonhos, precisamos que algumas
pessoas nos apoiem, e que outras nos dêem oportunidades
para realizá-los...*

*Dedico esta pesquisa a todos que me apoiaram, em
especial minha mãe, Helena, e a quem me deu essa
grande oportunidade, minha orientadora Sandra.*

*Ao professor Dr. Mauro Sigiz, que continuará
motivando muitas pesquisas...*

AGRADECIMENTOS

A Dra. Sandra Fabbri, por há um pouco mais de dois anos ter me selecionado para ser sua aluna e me dar a honra de tê-la como orientadora. Por lapidar meu conhecimento, direcionar minhas descobertas e sempre, sempre estar muito próxima, incentivando, questionando e sendo um exemplo de pessoa e profissional.

Aos meus pais, Eli e Helena, e minhas irmãs, Heloisa e Emanuela, pelo amor, apoio e preocupação. Por acreditarem sempre na minha capacidade e fazerem de tudo para que um dia, eu fosse mestre. Agradeço a eles por TUDO, e só eles sabem o quanto esse TUDO significa.

A minha avó Encarnação pela constante preocupação, ao meu avô Hernandes (*in memorian*), que sabiamente, sempre defendeu a “natureza” das pessoas, ao meu avô Montoro, pelas guloseimas e também minha avó Irene (*in memorian*), todos tão presentes na minha formação.

A toda minha família pelo incentivo, em especial aos meus tios Ângela e Aldo.

Ao Ivan Ervolino, por todo amor, companheirismo, amizade, carinho, cuidados, conselhos, compreensão e apoio incondicional, essenciais durante o último ano de mestrado.

A Carolina e a Thâmy, simplesmente por serem as melhores amigas-irmãs que alguém pode ter e por me darem forças em todos os momentos. Aos amigos de Macaubal, pela amizade e pelos momentos de descontração, essenciais para recarregar as energias durante o mestrado, em especial ao “Luizão” pela constante torcida.

Aos colegas mestrandos e alunos de Iniciação Científica do LaPES, em especial ao Daniel de Paula Porto, por ser um fantástico companheiro de mestrado e por toda a ajuda.

A todos da “Pastelaria do Ganso”, pelos momentos alegres que compartilharam comigo.

Aos amigos e professores da UNIFEV que torceram por mim, em especial ao professor Dr. Djalma Silva pelo incentivo. Estendo meus agradecimentos aos amigos da BW Informática, pela oportunidade, compreensão e influência profissional.

Aos amigos do mestrado, aos quais tive muita sorte em tê-los como companheiros nessa jornada repleta de sentimentos contraditórios, dificuldades e conquistas. Pessoas que em todos os momentos foram amáveis comigo e estarão sempre nas minhas lembranças. Um agradecimento especial a Ana Paula Santos, Kamila Rios, Matheus Viana, Ivan Zanon, Silviane Nascimento e Rodrigo Bella.

Ao meu primo Flávio Montoro por tornar nossa república minha casa, pelos desabafos e companheirismo, e também ao Renato Resina.

A toda a equipe do projeto Web-PIDE, em especial aos professores Marilde Terezinha Prado Santos e Ricardo Rodrigues Ciferri, e ao meu “fofo” companheiro de projeto, Thiago Luís Lopes Siqueira, pela simpatia e incentivo.

Aos professores e funcionários do Departamento de Computação da UFSCar.

A Deus, por colocar pessoas fantásticas no meu caminho.

A CAPES e ao INEP, pelo apoio financeiro por meio do Observatório da Educação.

*"Quem tem consciência para ter coragem...
Quem tem a força de saber que existe...
E no centro da própria engrenagem,
inventa contra a mola que resiste.*

*Que não vacila mesmo derrotado...
Quem já perdido nunca desespera...
E envolto em tempestade decepada,
entre os dentes segura a primavera."*

*Os Mutantes
João Ricardo/João Apolinário, 1973.*

RESUMO

Contexto: Há dezoito anos o INEP realiza censos e avaliações em instituições educacionais com o intuito de coletar dados que retratem a realidade educacional do Brasil. Entretanto, esse grande volume de dados encontra-se não uniformizado e distribuído em diferentes bases de dados, o que dificulta a realização de análises mais significativas. Para encontrar soluções para minimizar esses problemas, o INEP e a Capes vêm apoiando projetos de pesquisa, como o Web-PIDE, no qual este trabalho está inserido. A proposta desse projeto é criar uma plataforma para compartilhar e integrar esses dados, de forma que eles possam ser utilizados por toda a comunidade. **Objetivo:** Este trabalho teve dois objetivos. O primeiro foi criar uma Linguagem para Integração de Dados Educacionais, denominada LIDE, que dará suporte à implementação da plataforma proposta no projeto Web-PIDE. O segundo foi definir um processo para dar suporte à conceitualização de ontologias, uma vez que uma ontologia pode ser útil para o domínio das avaliações do INEP. **Metodologia:** Com base em visualização foi desenvolvida uma ferramenta e um processo para cada objetivo – a *SEV-Tool* e o processo P-LIDE (Processo para definição da Linguagem para Integração de Dados Educacionais) e a *ONTOP-Tool* e o processo ONTOP (*ONTOlogy Process*). **Resultados:** Foi criada a LIDE_{SAEB} que será usada na plataforma Web-PIDE e foi gerada uma lista de termos que compõe um glossário que dará suporte à criação de uma ontologia para as avaliações do INEP. **Conclusão:** O processo P-LIDE e a *SEV-Tool* compõem um apoio para o tratamento dos dados sobre as avaliações e pode ser aplicado a todas as bases de dados do INEP. O processo ONTOP e a *ONTOP-Tool* compõem um apoio para a conceitualização de ontologias, independentemente do domínio ao qual ela se refere. O uso da visualização como recurso para tratamento desses dados facilitou substancialmente a realização das tarefas propostas por ambos os processos.

Palavras-chave: tratamento de metadados, dados educacionais do INEP, ontologia, visualização.

ABSTRACT

Context: Aiming at collecting data that represent the educational reality of Brasil, from eighteen years the INEP applies census and evaluations on educational institutions. However, this large amount of data is non uniform and distributed among different databases, making difficult more significant analysis. Aiming to solve or minimize these problems, the INEP and Capes are supporting research projects, such as Web-PIDE project, in which this research was carried out. The objective of this project is to create a platform for sharing and integrating these data, which can be used by all the community. **Aim:** This work had two objectives. The first one was to create a Language for Integration of Educational Data, named LIDE that will support the Web-PIDE platform implementation. The second one was to define a process to support the ontology conceptualization phase, since an ontology can be useful for the INEP evaluation domain. **Methodology:** Based on visualization it was developed a tool and a process for each objective – the SEV-Tool and P-LIDE (Process to definition the Language for Integration Educational Data); and the ONTOP-Tool and ONTOP (*ONTO*logy *Process*). **Results:** It was created the LIDE|_{SAEB} that will be used in the platform Web-PIDE and it was generated a list of terms that composes a glossary that will support the creation of an ontology for INEP evaluations. **Conclusion:** The P-LIDE process and the SEV-Tool compose a support for the treatment of the INEP evaluation data. The ONTOP process and ONTO-Tool compose a support for ontology conceptualization, in spite of the domain. The use of visualization as the resource for the treatment of these data made easier the execution of the tasks proposed by both processes.

Keywords: metadata treatment, INEP educational data, ontology, visualization.

LISTA DE FIGURAS

Figura 2.1. Estrutura de Avaliações do INEP.	22
Figura 2.2. Tela de consulta da ferramenta Edudata Brasil para dados dos concluintes do Ensino Fundamental (EDUDATABRASIL, 2009).....	26
Figura 2.3. Tela de resultado apresentado em HTML para a consulta demonstrada na figura anterior (EDUDATABRASIL, 2009).	26
Figura 2.4. Tela com resultados de uma consulta feita sobre uma instituição de ensino do nível fundamental ao médio do interior do estado de São Paulo (DATAESCOLA, 2009)	27
Figura 2.5. Possíveis consultas sobre o Ensino Superior Brasileiro (SINAES, 2009).....	27
Figura 2.6. Possíveis consultas sobre o Ensino Superior Brasileiro (SINAES, 2009).....	28
Figura 2.7. Arquitetura da plataforma WebPIDE (TURINE <i>et al.</i> , 2006).....	30
Figura 2.8. Início de um dos arquivos de entrada SAS do SAEB 2003.....	32
Figura 2.9. Estrutura de um dos arquivos ASCII do SAEB 2003.....	32
Figura 2.10. Estrutura do arquivo “Leia-me.pdf” do SAEB 2003.	33
Figura 3.1. Reconhecimento de Padrões (NASCIMENTO; FERREIRA, 2005).....	37
Figura 3.2. (a) Gastos apresentados em tabela. (b) Gráfico de linhas. (c) Gráfico de pizza (NASCIMENTO; FERREIRA, 2005).	37
Figura 3.3. Processo simplificado de Visualização de Informações auxiliada por computador (NASCIMENTO; FERREIRA, 2005).	39
Figura 3.4. Modelo de Referência para Visualização (CARD <i>et al.</i> , 1999 <i>apud</i> NASCIMENTO; FERREIRA, 2005).....	39
Figura 3.5 – Exemplo de marcas visuais. Adaptado de Nascimento e Ferreira (2005).	40
Figura 3.6 - Propriedades gráficas das marcas. Adaptado de Nascimento e Ferreira (2005).	41
Figura 3.7. Exemplo da técnica Coordenadas Paralelas por Inselberg e Dimsdale (1990). .	43
Figura 3.8. Exemplo de utilização da técnica Coordenadas Paralelas (POUSMAN, STASKO, 2006).....	44
Figura 3.9. Faces de Chernoff (CHERNOFF, 1973).....	45
Figura 3.10. Tela da ferramenta PDBVis (Karve, Gleicher, 2007).....	45
Figura 3.11. Exemplo de utilização da técnica de Grafos (TOLLIS, 1996).....	46
Figura 3.12. Tela de um site que disponibiliza o resultado da busca do usuário por meio de grafos (LIVEPLASMA, 2009).	47
Figura 3.13. Proposta de Árvore Hiperbólica de Lamping e Rao (Lamping, Rao, 1996).	48

Figura 3.14. (a) <i>Browser</i> hiperbólico utilizado para visualizar um glossário de ESE. (b) Visualização dos termos que iniciam com a letra A no glossário de ESE (ESEGLOSSARY, 2009).....	48
Figura 3.15. (a) Visualização de código-fonte sem a técnica fish-eye. (b) Exemplificação de Furnas para técnica Fish-Eye (FURNAS, 1982).....	49
Figura 3.16. Menu <i>Fish-Eye</i> (FISHEYEMENU, 2009).....	50
Figura 3.17. Visualização da estrutura de diretórios de um disco rígido (JOHNSON, SHNEIDERMAN, 1991).	51
Figura 3.18 – Site de notícias que utiliza a técnica <i>Tree-Map</i> (NEWSMAP, 2009).....	51
Figura 3.19. Ferramenta TreeMap.....	52
Figura 3.20. Tela da ferramenta CRISTA (PORTO; MENDONÇA; FABBRI, 2008).....	53
Figura 3.21. Tela da ferramenta JFlap (JFLAP, 2009).	53
Figura 3.22. Tela da ferramenta Tulip (TULIP, 2009).....	54
Figura 4.1. Estágios de desenvolvimento de ontologias de Uschold e King (1995).....	62
Figura 4.2. O método Methontology. Adaptado de Gómez-Pérez, Fernández-López e Corcho (2003).	64
Figura 4.3. Relacionamento entre linguagens de ontologia baseada na Web. Adaptada de Gómez-Pérez, Fernández-López e Corcho (2003).....	69
Figura 5.1. Processo P-LIDE.	84
Figura 5.2. Tela de boas vindas da SEV- <i>Tool</i>	86
Figura 5.3. Tela inicial da SEV- <i>Tool</i>	87
Figura 5.4. Tela com as funcionalidades da SEV- <i>Tool</i>	88
Figura 5.5. Formato de entrada para a visualização dos questionários na SEV- <i>Tool</i>	89
Figura 5.6. Tela da SEV- <i>Tool</i> para visualização dos questionários organizados por nome.	91
Figura 5.7. Tela da SEV- <i>Tool</i> para visualização dos questionários organizados por ano. ...	92
Figura 5.8. Tela da ONTOP- <i>Tool</i> para geração da LIDE.	93
Figura 5.9. Tela da SEV- <i>Tool</i> para consultas na LIDE.....	94
Figura 5.10. O processo ONTOP.	95
Figura 5.11. Tela de boas vindas da ONTOP- <i>Tool</i>	98
Figura 5.12. Tela inicial da ONTOP- <i>Tool</i>	98
Figura 5.13. Tela com as funcionalidades da ONTOP- <i>Tool</i>	99
Figura 5.14. Tela de glossário no Moodle.	100
Figura 5.15. Tela da ONTOP- <i>Tool</i> para visualização do glossário.	101
Figura 5.16. Tela da ONTOP- <i>Tool</i> para classificação dos termos.	104
Figura 5.17. Tela da ONTOP- <i>Tool</i> para hierarquização das classes.	105
Figura 5.18. Tela da ONTOP- <i>Tool</i> para hierarquização das classes.	107

Figura 6.1. Tela de visualização dos questionários de português do SAEB.	116
Figura 6.2. Tela da ONTOP- <i>Tool</i> no início da padronização sintática.	118
Figura 6.3. Tela da ONTOP- <i>Tool</i> após a padronização de cinco questões.	118
Figura 6.4. Tela da ONTOP- <i>Tool</i> após a padronização de todas as questões.	119
Figura 6.5. Tela para a geração da LIDE.	122
Figura 6.6. Organização da LIDE.	122
Figura 6.7. Tela do Moodle com o glossário gerado com apoio da ONTOP- <i>Tool</i>	123
Figura 6.8. Tela da visualização inicial do glossário de ESE na ONTOP- <i>Tool</i>	125
Figura 6.9. Tela da visualização do glossário de ESE na ONTOP- <i>Tool</i> após a adequação ao contexto de Experimentos Controlados.	126
Figura 6.10. Tela da ONTOP- <i>Tool</i> para definir componentes da ontologia.	127
Figura 6.11. Tela da ONTOP- <i>Tool</i> para definição de hierarquia.	128
Figura 6.12. Tela da ONTOP- <i>Tool</i> para definir <i>Domain-Range</i> das propriedades.	130
Figura 6.13. Tela da Protégé-2000 com as classes instanciadas.	131
Figura 6.14. Tela da Protégé-2000 com as propriedades instanciadas.	131
Figura 6.15. Visualização do arquivo OWL por meio do plug-in Jambalaya (2009).	132

LISTA DE ABREVIATURAS E SIGLAS

AAD	Ambiente de Apoio à Decisão
ADW	Ambientes de Data Warehouse
ASCII	<i>American Standard Code for Information Interchange</i>
BioML	<i>Biopolymer Markup Language</i>
CML	<i>Chemical Markup Language</i>
CSS	<i>Cascading Style Sheets</i>
DAML	<i>Darpa Agent Markup :Language</i>
DM	<i>Data Marts</i>
DW	<i>Data Warehouse</i>
Enade	Exame Nacional de Desempenho de Estudantes
ENC	Exame Nacional de Cursos
Enem	Exame Nacional do Ensino Médio
GLEO	<i>Graphical Language for Expressing Ontologies</i>
GML	Generalized Markup Language
HTML	<i>HyperText Markup Language</i>
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LIDE	Linguagem para Integração de Dados Educacionais
MathML	<i>Mathematical Markup Language</i>
OCML	<i>Operational Conceptual Modeling Language</i>
OIL	<i>Ontology Interchange Language</i>
OKBC	<i>Open Knowledge Base Connectivity</i>
OWL	<i>Ontology Web Language</i>
OWLApi	<i>Ontology Web Language Application Programming Interface</i>
PDF	<i>Portable Document Format</i>
RDF	<i>Resource Description Framework</i>
SAD	Sistemas de Apoio a Decisão
Saeb	Sistema Nacional de Avaliação da Educação Básica
SAS	<i>Statistical Analysis Software,</i>
SGBD	Sistema de Gerenciamento de Banco de Dados
SGI	Sistemas de Gestão de Informação
SHOE	<i>Simple HTML Ontology Extension</i>
STMML	<i>Scientific, Technical and Medical Markup Language</i>
W3C	<i>World Wide Web Consortium</i>

XACDML	<i>eXtensible Activity Cycle Diagrama Markup Language</i>
XML	<i>Extensible Markup Language</i>
XOL	<i>XML Ontology Language</i>
XrML	<i>eXtensible rights Markup Language</i>

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO	16
1.1 Contexto	16
1.2 Motivação e objetivos.....	18
1.3 Metodologia de desenvolvimento do trabalho	18
1.4 Organização do trabalho.....	19
CAPÍTULO 2 - O PROJETO WEB-PIDE	21
2.1 Considerações Iniciais	21
2.2 As avaliações do INEP.....	22
2.2.1 Divulgação dos resultados	25
2.3 O projeto Web-PIDE	29
2.3.1 O formato dos dados disponibilizados para o projeto.....	31
2.4 Considerações finais	34
CAPÍTULO 3 - VISUALIZAÇÃO DE INFORMAÇÕES	35
3.1 Considerações Iniciais	35
3.2 Visualização de Informações	36
3.3 Técnicas de Visualização.....	42
3.3.1 Coordenadas Paralelas	43
3.3.2 Glifos	44
3.3.3 Grafos.....	46
3.3.4 Árvore Hiperbólica.....	47
3.3.5 Fish-Eye	49
3.3.6 <i>Tree-Map</i>	50
3.4 Ferramentas de Visualização.....	52
3.5 Trabalhos relacionados.....	55
3.6 Considerações Finais.....	57
CAPÍTULO 4 - ONTOLOGIAS	59
4.1 Considerações Iniciais	59
4.2 Ontologias	60

4.3 Desenvolvimento de ontologias	62
4.3.1 O método Methontology	64
4.4 Ferramentas e linguagens para construção de ontologias.....	68
4.4.1 Características da OWL	71
4.5 Trabalhos relacionados.....	75
4.6 Considerações Finais.....	79
CAPÍTULO 5 - OS PROCESSOS P-LIDE E ONTOP E AS FERRAMENTAS DE APOIO.....	81
5.1 Considerações Iniciais	81
5.2 O P-LIDE e a SEV- <i>Tool</i>	83
5.3 A SEV- <i>Tool</i> e suas funcionalidades.....	86
5.4 O processo ONTOP e a ONTOP- <i>Tool</i>	94
5.5 ONTOP- <i>Tool</i> e suas funcionalidades	97
5.6 Considerações Finais.....	108
CAPÍTULO 6 - APLICAÇÃO DOS PROCESSOS P-LIDE E ONTOP	111
6.1 Considerações Iniciais	111
6.2 Processo P-LIDE: um exemplo com os dados do INEP.....	112
6.2.1 Passo 1: Preparar os dados do SAEB para serem importados na SEV- <i>Tool</i>	112
6.2.2 Passo 2: Gerar a LIDE	115
6.3 Processo ONTOP: um exemplo com o glossário de Engenharia de Software Experimental	124
6.3.1 Passo 1: refinar o glossário de ESE.....	124
6.3.2 Passo 2: definir componentes da ontologia de Experimentos Controlados....	126
6.3.3 Passo 3: definir hierarquia das classes	127
6.3.4 Passo 4: definir relacionamento entre as classes	128
6.3.5 Passo 5: gerar arquivo OWL	130
6.4 Considerações Finais.....	132
CAPÍTULO 7 - CONCLUSÕES.....	134
7.1 Contribuições e limitações	136
7.2 Trabalhos futuros	139
REFERÊNCIAS.....	140

APÊNDICE A - COMPARAÇÕES ENTRE ARQUIVOS DO SAEB149

APÊNDICE B - OBSERVAÇÕES SOBRE OS QUESTIONÁRIOS DO SAEB154

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto do projeto Web-PIDE, no qual esta pesquisa se insere, expondo a motivação, os objetivos, a metodologia utilizada para concretizá-la e a organização desta dissertação.

1.1 Contexto

Com o objetivo de produzir dados que indiquem a realidade do setor educacional do Brasil e auxilie a determinação de políticas para melhoria da educação, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), há dezoito anos, realiza censos e avaliações nas instituições de ensino em todo o país.

No entanto, todos esses dados acumulados pelo INEP encontram-se não uniformizadas e distribuídas em diferentes bases de dados, dificultando a realização de análises, informações estatístico-educacionais e dificultando o reúso e possíveis integrações com outros sistemas.

Com o objetivo de minimizar esse problema, em 2006, por meio do programa Observatório da Educação, criado pelo INEP em parceria com a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), foi aprovado o projeto Web-PIDE, com parceiros da UFSCar e da UFMS, no contexto do qual esta pesquisa está inserida. O objetivo desse projeto é especificar e implementar uma plataforma para compartilhar e integrar os dados educacionais do INEP, visando facilitar a consulta e geração de hipóteses relativas a esses dados e minimizar o problema com as bases de dados do INEP.

Um dos pilares para a construção da plataforma é a LIDE – Linguagem para Integração de Dados Educacionais, objeto inicial desta pesquisa, que deve explicitar, dentre as questões que compõem os questionários dos censos e avaliações os anos e os questionários em que cada questão foi utilizada, o atributo correspondente da base de dados do INEP e o tipo desse atributo. Assim, a função da LIDE é servir de intermediária entre o banco de dados e a plataforma.

Embora no início parecesse uma tarefa relativamente simples, após um primeiro contato com os dados das avaliações pode-se perceber que para analisar todas as bases de dados das diversas avaliações seria fundamental um apoio computacional. Por outro lado, esse apoio deveria ser tal que facilitasse a análise humana, pois não seria possível automatizá-lo por completo, uma vez que uma mera comparação de duas questões poderia colocá-las como distintas pela diferença de um acento gráfico. Além dessa situação, outro ponto que logo se mostrou um problema foi que os termos utilizados na formulação das questões não tinham o mesmo significado em todos os momentos que eram usados, mostrando a necessidade de se ter um vocabulário único, compartilhado por todos os questionários, sugerindo a criação de uma ontologia para o domínio de avaliações educacionais do INEP.

Dessa forma, a necessidade de definir a LIDE para dar suporte à plataforma proposta no projeto Web-PIDE incentivou a construção de um processo sistematizado, apoiado computacionalmente, para facilitar a execução dessa tarefa, atentando-se para os problemas indicados anteriormente.

Além disso, o objetivo do projeto Web-PIDE de compartilhar um grande volume de dados e informações por meio de tecnologias web tem sido, atualmente, uma necessidade constante. De acordo com Daconta, Obrst e Smith (2003), mecanismos apropriados para essa tarefa constituem alvo de pesquisa no contexto de Web Semântica, sendo que um recurso de auxílio para essa área são as ontologias, o que reforçou o quão útil seria a criação de uma ontologia para o domínio das avaliações educacionais do INEP.

1.2 Motivação e objetivos

Dado o problema apresentado na seção anterior, a motivação desta pesquisa foi desenvolver um processo sistematizado, com apoio computacional, para possibilitar a definição da LIDE.

Como era necessário analisar todos os questionários a fim de identificar as avaliações em que cada questão foi utilizada e considerando que muitas vezes a mesma questão não era escrita com a mesma sintaxe e que termos eram utilizados com significados diferentes, ficou evidente que o apoio computacional deveria ser parcial, com o objetivo de facilitar a decisão humana e não de automatizar o processo por completo.

Assim, decidiu-se utilizar uma técnica de visualização para permitir que o usuário analisasse todos os dados e, além de realizar a tarefa de padronização sintática das questões, pudesse abstrair rapidamente novas informações sobre os diversos questionários e conhecer melhor a estrutura das avaliações.

Considerando que um apoio baseado em técnica de visualização poderia ajudar efetivamente esse primeiro problema, decidiu-se explorar a mesma técnica para apoiar a definição de uma ontologia para esse contexto de avaliações do INEP, com o objetivo de padronizar a utilização dos termos usados nas questões.

1.3 Metodologia de desenvolvimento do trabalho

Os passos dos processos P-LIDE (Processo para geração da LIDE) e ONTOP (ONTOlogy conceptualization Process), propostos nesta dissertação, foram norteados pelas dificuldades e necessidades da pesquisa. Da mesma forma ocorreu o desenvolvimento dos suportes computacionais desenvolvidos, denominados *SEV-Tool* e *ONTOP-Tool*, que tiveram suas funcionalidades implementadas à medida que os passos dos dois processos foram definidos.

A primeira etapa realizada foi a análise e padronização sintática das questões presentes nas avaliações do INEP para a geração da LIDE. Para isso, aplicou-se o recurso de visualização baseado na técnica *Tree-Map* para representar as questões,

agrupá-las, prover recursos de busca e de edição. Com isso, as questões foram padronizadas tornando possível a geração da LIDE.

Em uma segunda etapa, depois de resolvido o problema de geração da LIDE, criaram-se passos para dar suporte à conceitualização da ontologia para as avaliações educacionais do INEP. Para apoiar a solução desse segundo problema manteve-se o suporte provido pelos recursos de visualização e utilizou-se também o recurso de construção de glossários provido pelo ambiente Web de aprendizado Moodle, principalmente pelo fato dele proporcionar uma participação coletiva e distribuída. Nessa etapa, a abordagem proposta procura facilitar algumas definições iniciais de uma ontologia. Embora tenha sido norteadada pelo problema do INEP, o processo proposto é genérico e pode ser aplicado como suporte para a definição de qualquer ontologia.

Como a definição de uma ontologia é complexa e exige um grande conhecimento sobre o domínio, apoiou-se no método Methontology, considerado um dos mais maduros de acordo com a literatura. Constatando-se que a fase de Conceitualização do método é uma das mais importantes e que não se encontrou ferramentas computacionais que apoiassem sua execução, essa etapa do processo proposto procurou dar suporte computacional a ela, mantendo os recursos de visualização adotados na primeira etapa. Assim, por meio de um processo iterativo um glossário de termos do domínio em questão é refinado por especialistas, alternando-se as informações entre o ambiente Moodle e a *ONTOP-Tool*. Após a definição dos elementos da ontologia com o apoio da visualização, é possível gerar um arquivo OWL com as definições estabelecidas, de forma que ele possa ser usado por outras ferramentas para que a construção da ontologia seja finalizada.

1.4 Organização do trabalho

Esta dissertação está organizada em 8 Capítulos e 2 Apêndices, sendo que seus respectivos objetivos são resumidos a seguir.

O Capítulo 1 – Introdução, caracteriza o contexto e evidencia as lacunas que motivaram o desenvolvimento desta pesquisa, apresentando também seus objetivos.

O Capítulo 2 – O projeto Web-PIDE, apresenta as avaliações do INEP, as maneiras como os resultados são divulgados, os dados que o INEP disponibilizou para o projeto e descreve o projeto Web-PIDE.

O Capítulo 3 – Visualização de Informações, apresenta conceitos sobre visualização, ilustra algumas técnicas e exemplifica algumas ferramentas. Alguns trabalhos relacionados também são mencionados nesse capítulo.

O Capítulo 4 – Ontologia, apresenta conceitos sobre ontologia, métodos para sua construção, dando ênfase ao método Methontology, ferramentas e linguagens para construção de ontologias, enfatizando a linguagem OWL. Assim como no Capítulo 3, alguns trabalhos relacionados também são mencionados nesse capítulo.

O Capítulo 5 – Os processos P-LIDE e ONTOP e as ferramentas de apoio, apresenta detalhadamente os processos e as ferramentas resultantes dessa pesquisa.

O Capítulo 6 – Aplicação dos processos P-LIDE e ONTOP, apresenta a aplicação de cada processo, apoiado respectivamente pela *SEV-Tool* e pela *ONTOP-Tool*, sendo que uma aplicação é feita com dados do INEP e a outra, com um glossário de Engenharia de Software Experimental.

O Capítulo 7 – Conclusões, apresenta as conclusões obtidas na pesquisa, as limitações detectadas e sugestões de trabalhos futuros.

O Apêndice A apresenta os problemas encontrados entre os arquivos disponibilizados pelo INEP para o projeto Web-PIDE e o Apêndice B apresenta um relatório sobre questões consideradas confusas ou que contêm termos usados com significados distintos.

Capítulo 2

O PROJETO WEB-PIDE

Este capítulo descreve o projeto Web-PIDE – Plataforma para Integração de Dados Educacionais na Web - que foi a motivação para esta pesquisa.

2.1 Considerações Iniciais

A produção de dados e informações estatístico-educacionais de forma ágil e fiel, que retrate a realidade do setor educacional, é um dos instrumentos básicos de avaliação, planejamento e auxílio ao processo decisório para o estabelecimento de políticas de melhoria da educação brasileira.

Nesse contexto, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) estruturou os Sistemas Nacionais de Avaliação e de Informação, criando diversos censos, avaliações e pesquisas, com o objetivo de avaliar e acompanhar a evolução do ensino no país em todos os âmbitos: ensino regular (básico e fundamental), médio e superior.

Com o intuito de estimular a utilização dos dados adquiridos em pesquisas acadêmicas, contribuindo também para a formação de recursos humanos pós-graduados (mestrado e doutorado), o INEP, em parceria com a CAPES, criou o programa Observatório da Educação.

O Observatório da Educação prevê que as pesquisas utilizem os dados das avaliações, censos e pesquisas do INEP como subsídio para estudos sobre a realidade educacional do país. O primeiro edital deste programa foi lançado em 2006, quando foi submetido e aprovado o projeto Web-PIDE (INEP, 2009). Em 2008 um novo edital foi lançado.

Este Capítulo apresenta na Seção 2.2 as avaliações do INEP e como é feita a divulgação dos resultados e na Seção 2.3, o projeto Web-PIDE e os dados disponibilizados pelo INEP para o projeto.

2.2 As avaliações do INEP

A Figura 2.1 ilustra a estrutura das avaliações criada pelo INEP e, a seguir, é apresentada uma breve descrição das mesmas.

DOMÍNIO	Ensino Regular					Ensino Médio	Ensino Superior				
SISTEMA DE AVALIAÇÃO			SAEB Sistema Nacional de Avaliação da Educação Básica					SINAES Sistema Nacional de Avaliação da Educação Superior			
AVALIAÇÃO	CE ↓ Censo Escolar	ENCCEJA ↓ Exame Nacional para Certificação de Competências de Jovens e Adultos	Aneb ↓ Avaliação Nacional da Educação Básica	Ansresc ↓ Avaliação Nacional do Rendimento Escolar	Provinha Brasil ↓ Avaliação da Alfabetização	Enem ↓ Exame Nacional do Ensino Médio	ENC ↓ Exame Nacional de Cursos	Censo da Educação Superior	Avaliação das Instituições de Educação Superior	Avaliação dos Cursos de Graduação	Enade ↓ Exame Nacional de Desempenho de Estudantes
NOME DE DIVULGAÇÃO	Censo Escolar	ENCCEJA	Saeb	Prova Brasil	Provinha Brasil	Enem	Provão	Censo da Educação Superior	Avaliação das Instituições de Educ. Superior	Avaliação dos Cursos de Graduação	Enade
ANO DE APLICAÇÃO	1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008	2002 2003 2004 2005 2006 2007 2008	1991 1993 1995 1997 1999 2001 2003 2005 2007	2005 2007	1º Bim 2008 2º Bim 2008 1º Bim 2009	1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008	1996 1997 1998 1999 2000 2001 2002 2003	1996 1997 1998 1999 2000 2001 2002 2003	2002 2003 2004 2005 2006 2007 2008 2009	2004 2005 2006 2007 2008 2009	2004 2005 2006 2007 2008

Figura 2.1. Estrutura de Avaliações do INEP.

De acordo com o INEP (2009):

Censo Escolar: é um levantamento de dados estatístico-educacionais de âmbito nacional realizado todos os anos. Ele é feito com a colaboração das secretarias estaduais e municipais de Educação e com a participação de todas as escolas públicas e privadas do país.

É considerado o principal instrumento de coleta de informações da educação básica, que abrange o ensino regular (educação infantil e ensinos fundamental e médio), educação especial e educação de jovens e adultos (EJA). Neste censo são coletados dados sobre os estabelecimentos de ensino, matrículas, funções dos docentes, movimento e rendimento escolar.

ENCCEJA: o Exame Nacional para Certificação de Competências de Jovens e Adultos é uma avaliação que mede as competências e habilidades de jovens e adultos, residentes no Brasil e no exterior, que estão concluindo o Ensino Fundamental e Médio. A adesão a esta avaliação é opcional e realizada pelas secretarias de Educação dos Estados, Distrito Federal e Municípios.

SAEB: foi a primeira iniciativa brasileira, em escala nacional, para se conhecer o sistema educacional brasileiro em profundidade. É composto por dois processos: a Avaliação Nacional da Educação Básica (Aneb) e a Avaliação Nacional do Rendimento Escolar (Anresc).

Aneb: teve sua primeira edição em 1991 e é realizada por amostragem das Redes de Ensino, em cada unidade da Federação e tem foco nas gestões dos sistemas educacionais. Por manter as mesmas características, a Aneb recebe o nome do Saeb em suas divulgações;

Anresc: criada em 2005 é mais extensa e detalhada que a Aneb e tem foco em cada unidade escolar. Por seu caráter universal, é divulgada com o nome de Prova Brasil. As duas avaliações acontecem simultaneamente e são aplicadas de dois em dois anos.

Provinha Brasil: esta avaliação é uma iniciativa inaugurada pelo Ministério da Educação (MEC) no primeiro semestre de 2008, com o objetivo de oferecer aos professores, diretores, coordenadores e gestores das redes públicas de ensino um instrumento de diagnóstico do nível de alfabetização das crianças com idade entre seis e oito anos de idade. O INEP/MEC disponibiliza, anualmente, duas versões da Provinha Brasil. A primeira no 1º Bimestre do ano letivo e, a segunda, no 4º semestre do ano letivo, caracterizando, assim, o ciclo da prova.

Enem: é um exame individual, de caráter voluntário, oferecido anualmente desde 1998 aos estudantes que estão concluindo ou que já concluíram o ensino médio em anos anteriores. Seu principal objetivo é possibilitar uma referência para auto-avaliação, a partir das competências e habilidades que estruturam o exame.

ENC: conhecido como Provão, o Exame Nacional de Cursos foi aplicado aos formandos das instituições de ensino superior no período de 1996 a 2003. O objetivo

era avaliar os cursos de graduação no que tange aos resultados do processo de ensino-aprendizagem. Em 2004 o Provão foi substituído pelo Exame Nacional de Desempenho dos Estudantes (Enade).

SINAES: criado em 2004, o Sistema Nacional de Avaliação da Educação Superior é formado por três componentes principais: a Avaliação das Instituições de Ensino Superior, a Avaliação dos Cursos de Graduação e o Exame Nacional de Desempenho dos Estudantes.

O Sinaes avalia todos os aspectos em torno desses três eixos: o ensino, a pesquisa, a extensão, a responsabilidade social, o desempenho dos alunos, a gestão da instituição, o corpo docente, as instalações e outros.

Avaliação das Instituições de Ensino Superior: dividido em duas modalidades: a **Auto-avaliação** - coordenada pela Comissão Própria de Avaliação (CPA) de cada instituição e orientada pelas diretrizes e por um roteiro da auto-avaliação institucional; e a **Avaliação externa** – realizada por comissões designadas pelo INEP, que tem como referência os padrões de qualidade para a educação superior expressos nos instrumentos de avaliação e os relatórios das auto-avaliações.

Avaliação dos Cursos de Graduação: o INEP/MEC prevê que os cursos sejam avaliados periodicamente. Assim, os cursos de educação superior passam por três tipos de avaliação: para autorização, para reconhecimento e para renovação de reconhecimento. As avaliações dos cursos de graduação produzem indicadores que subsidiam tanto o processo de regulamentação, exercido pelo MEC, como garante transparência dos dados sobre qualidade da educação superior.

ENADE: aplicado pela primeira vez em 2004, o exame tem o objetivo de aferir o rendimento dos alunos dos cursos de graduação em relação aos conteúdos programáticos, suas habilidades e competências. O exame é realizado por amostragem e a participação ou, quando for o caso, sua dispensa pelo MEC, consta no histórico escolar do estudante. O INEP/MEC constitui a amostra dos participantes a partir da inscrição, pela própria instituição de ensino superior, dos alunos habilitados a fazer a prova.

Além destas avaliações, o INEP realiza diversos outros censos e pesquisas. Como exemplo destes censos podem-se citar o Censo Escolar Indígena, realizado

em 1999, Censo da Educação Infantil em 2000 e o Censo dos Profissionais do Magistério da Educação Básica em 2003.

Por não serem realizados periodicamente, estes, assim como outros censos e pesquisas, não são abordados no Projeto Web-PIDE, que será apresentado na Seção 2.2 deste capítulo e, conseqüentemente, não são utilizados nesta pesquisa.

Neste capítulo é apresentada a forma de divulgação e disponibilização desses dados pelo INEP e o projeto Web-PIDE, que visa amenizar as dificuldades encontradas atualmente nesse contexto.

2.2.1 Divulgação dos resultados

Para divulgar os resultados obtidos nas avaliações realizadas com os alunos, o INEP dispõe de três ferramentas: Edudata Brasil (EDUDATABRASIL, 2009), Data Escola Brasil (DATAESCOLABRASIL, 2009) e Portal Sinaes (SINAES, 2009).

O Sistema de Estatísticas Educacionais (Edudata Brasil) é um instrumento de divulgação dos dados educacionais tratados pelo INEP. Por meio dele, pesquisadores em educação, gestores e os demais interessados podem acompanhar melhor a evolução da educação no país. Apesar da sua interface facilitar a consulta por possuir um amplo leque de alternativas para o acesso ao banco de dados do INEP, como se pode ver na Figura 2.2, os resultados são apresentados em forma de tabela HTML (*Hyper Text Markup Language*), como ilustra a Figura 2.3. Essa limitação impede que o usuário manipule os dados resultantes da sua consulta. Recentemente foi disponibilizada a funcionalidade para visualizar os dados em planilha eletrônica.

Na Figura 2.3 é possível observar que o número de concluintes do Ensino Fundamental esteve em crescimento de 1998 até 2002 e que em 2003 esse número diminuiu, assim como em 2004. Em 2005 ocorreu um aumento em relação aos últimos dois anos, porém, como os dados da ferramenta não estão atualizados, atualmente não é possível saber se essa tendência permaneceu.

O sistema de consulta ao banco de dados da Educação Básica, Data Escola Brasil, permite o acesso às informações sobre as instituições do nível creche ao ensino médio, do perímetro urbano e rural, nos âmbitos federal, estadual, municipal ou privado. O sistema fornece informações como estatísticas básicas das escolas, equipamentos em uso, dependências existentes, infraestrutura disponível,

indicadores educacionais e programas dos quais a escola participa e participou. Da mesma maneira que a ferramenta citada anteriormente, os dados são apresentados em formato de tabelas, conforme pode ser observado na Figura 2.4, o que gera a mesma limitação que impede que o usuário manipule os dados resultantes da sua consulta.



Figura 2.2. Tela de consulta da ferramenta Edudata Brasil para dados dos concluintes do Ensino Fundamental (EDUDATABRASIL, 2009).

Abrangência Geográfica	Ano	Concluintes - total
Brasil	1.998	2.383.207
Brasil	1.999	2.484.972
Brasil	2.000	2.648.638
Brasil	2.001	2.707.683
Brasil	2.002	2.778.033
Brasil	2.003	2.668.605
Brasil	2.004	2.462.319
Brasil	2.005	2.471.690

Fonte: MEC/INEP

Figura 2.3. Tela de resultado apresentado em HTML para a consulta demonstrada na figura anterior (EDUDATABRASIL, 2009).

O Portal Sinaes permite consultar dados do Ensino Superior. A ferramenta oferece três componentes principais de consulta, através dos quais é permitido conhecer o perfil de determinada instituição de ensino ou de algum curso de graduação, obter informações sobre a Educação Superior no Brasil ou então visualizar dados sobre docentes, cursos e instituições.

Apesar das informações serem apresentadas em tabelas e/ou gráficos, o que facilita a interpretação dos dados, as consultas são estáticas, não permitindo que os

usuários as adéquem de acordo com suas necessidades. A Figura 2.5 ilustra a tela com as possíveis consultas sobre o Ensino Superior no país, enquanto a Figura 2.6 apresenta o resultado do total de concluintes do Ensino Superior de 1991 até 2005, separados por região do país.

The screenshot shows the 'Data Escola Brasil 2009' interface. It features a navigation bar with 'VOLTAR' and tabs for 'Dados Cadastrais', 'Ensino Regular', 'Educação Especial', 'Educação de Jovens e Adultos', and 'Cursos de educação profissional oferecidos'. The main content area displays two tables: 'ENSINO FUNDAMENTAL' and 'ENSINO MÉDIO GERAL'. Each table has columns for 'ETAPA', 'TURMAS', 'MATRÍCULA', and 'FUNÇÕES DOCENTES¹'. The 'ENSINO FUNDAMENTAL' table shows data for 'TOTAL', 'Com menos de 6 horas de duração', 'Com 6 horas de duração ou mais', and 'Com atendimento complementar²,¹¹'. The 'ENSINO MÉDIO GERAL' table shows data for 'TOTAL', 'Com menos de 6 horas de duração', and 'Com 6 horas de duração ou mais'.

ETAPA	TURMAS	MATRÍCULA	FUNÇÕES DOCENTES¹
ENSINO FUNDAMENTAL			
TOTAL	12	376	97
Com menos de 6 horas de duração	12	376	97
Com 6 horas de duração ou mais	0	0	0
Com atendimento complementar²,¹¹	0	0	0
Anos Iniciais³			
TOTAL	0	0	0
Com menos de 6 horas de duração			
Com 6 horas de duração ou mais			
Com atendimento complementar²			
Anos Finais⁸			
TOTAL	12	376	97
Com menos de 6 horas de duração	12	376	97
Com 6 horas de duração ou mais			
Com atendimento complementar²			
Multi⁵	0		
Correção de fluxo⁶	0		
Multi 8 e 9 anos⁹			
ENSINO MÉDIO GERAL			
TOTAL	9	291	89
Com menos de 6 horas de duração	9	291	89
Com 6 horas de duração ou mais	0	0	0

Figura 2.4. Tela com resultados de uma consulta feita sobre uma instituição de ensino do nível fundamental ao médio do interior do estado de São Paulo (DATAESCOLA, 2009)

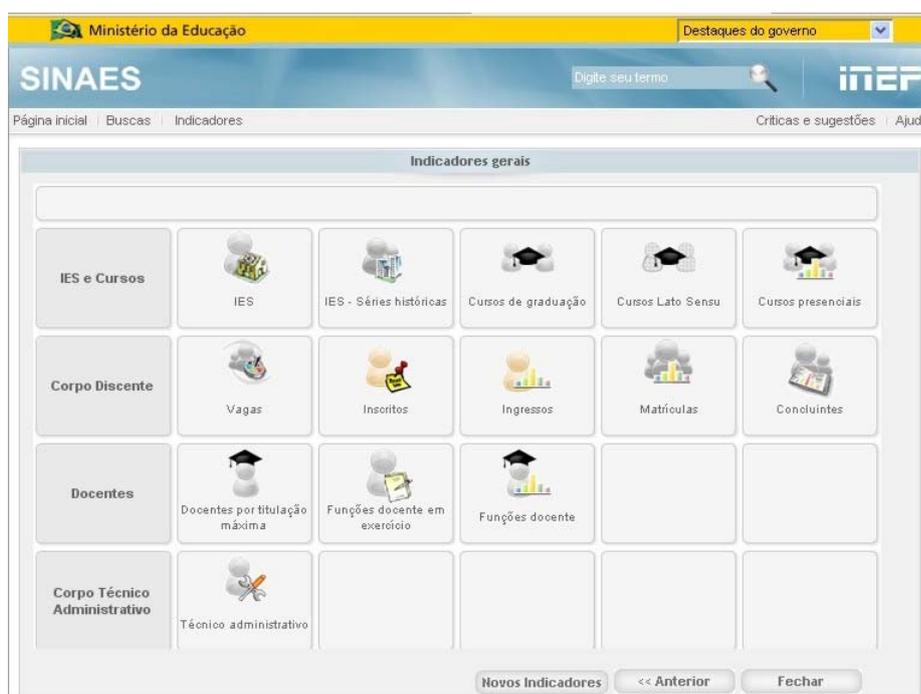


Figura 2.5. Possíveis consultas sobre o Ensino Superior Brasileiro (SINAES, 2009).

É possível notar, ao observar o gráfico da Figura 2.6, a discrepância que há entre o número de graduados por ano na região sudeste em relação às outras regiões. Nota-se também que em 2004 houve um aumento no número de graduados na região norte.

Nota-se que, assim como nas outras ferramentas utilizadas pelo INEP, a atualização dos dados é lenta, não permitindo ao usuário ter acesso a informações recentes, além de prejudicar a consulta, pois dependendo da avaliação consultada e da informação que o usuário busca, o resultado pode não retratar o cenário atual.



Figura 2.6. Possíveis consultas sobre o Ensino Superior Brasileiro (SINAES, 2009).

Além das ferramentas citadas acima, o INEP disponibiliza, por meio do seu site (<http://www.inep.gov.br>), uma maneira para que os microdados (arquivo em formato texto e o dicionário de variáveis) de algumas avaliações e censos sejam acessados. É possível requisitá-los via e-mail e receber um CD-ROM em um endereço especificado, ou adquiri-los via download, diretamente no site.

Ao analisar as ferramentas disponibilizadas pelo INEP, é possível constatar que a maneira disponibilizada para consultas aos dados obtidos nas avaliações não

é adequada para gestores da educação, para pais, pesquisadores e outros que possam se interessar por essas informações. O fácil acesso aos microdados compartilha os dados com muitos, porém apenas alguns grupos de pessoas conseguem manipulá-los, levando em consideração que são necessários conhecimentos específicos para essa tarefa.

Considerando as limitações e dificuldades citadas nesta Seção, desde 2006 o INEP, em parceria com CAPES, promove chamadas de projetos para o Observatório da Educação. Nesse contexto foi submetido e aprovado em 2006 o projeto Web-PIDE, apresentado na seção seguinte.

2.3 O projeto Web-PIDE

Em dezoito anos de censos e pesquisas, várias bases de dados foram produzidas para possibilitar o armazenamento e análise das informações adquiridas.

Com a integração desses dados é possível traçar um quadro abrangente da situação educacional do país e fornecer subsídios indispensáveis para o aprofundamento de análises e pesquisas críticas que possam enriquecer o debate sobre os rumos da educação brasileira.

No entanto, essas informações encontram-se distribuídas em diferentes bases de dados e de maneira não uniformizada, fato que dificulta seu reuso e possíveis integrações com sistemas de avaliação institucional das universidades nacionais (TURINE *et al.*, 2006).

Atualmente, o cruzamento de informações do SAEB, que avalia o ensino básico, e do ENEM, que avalia o ensino médio, só é possível manualmente ou através de algum ferramental desenvolvido pelo interessado.

O fato das bases de dados armazenarem um grande volume de informações também dificulta sua análise e impossibilita uma tomada de decisão eficiente baseada nos dados das avaliações.

Neste escopo, o projeto Web-PIDE tem o objetivo de especificar e implementar uma plataforma para compartilhar e integrar dados educacionais, visando facilitar a consulta aos dados das avaliações e geração de hipóteses relativas aos dados armazenados. A Figura 2.7 apresenta a arquitetura da plataforma.

Para integrar os dados educacionais, o projeto propôs uma linguagem de marcação intitulada LIDE (Linguagem para Integração de Dados Educacionais) para cada uma das bases de dados do INEP. As LIDEs são definidas com a linguagem de marcação XML (*Extensible Markup Language*), que torna os documentos legíveis para as pessoas e manipuláveis pelos computadores.

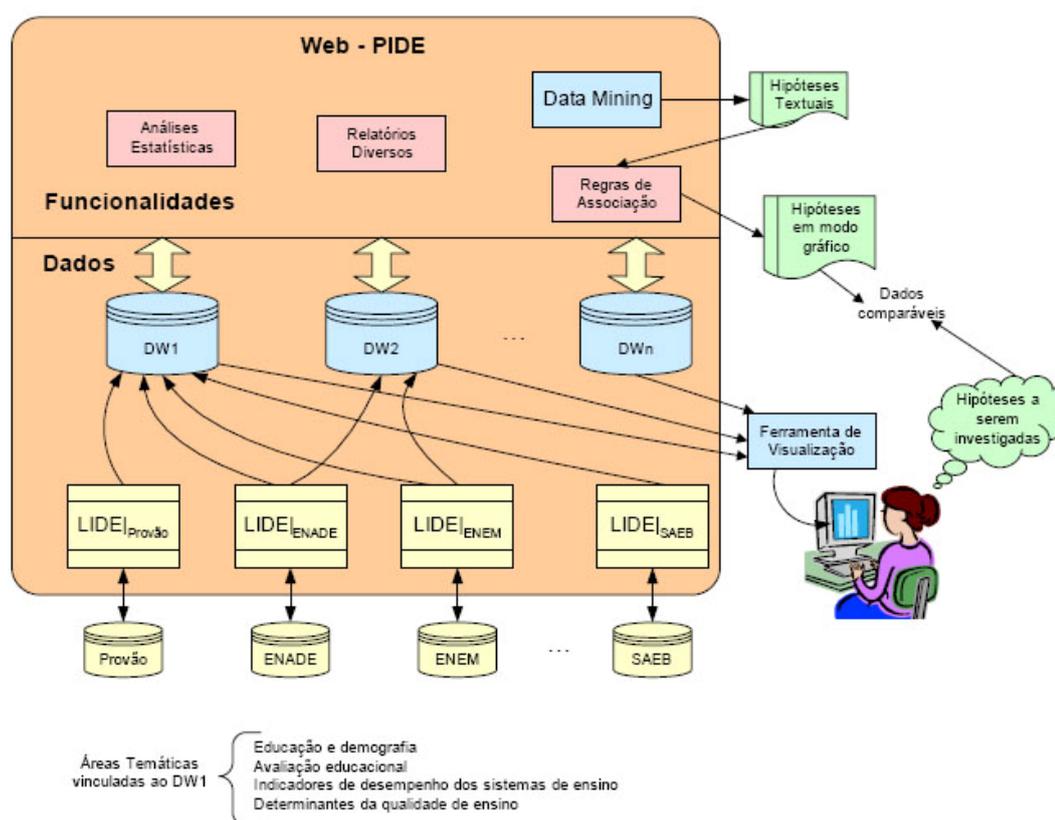


Figura 2.7. Arquitetura da plataforma WebPIDE (TURINE *et al.*, 2006).

A LIDE deve possuir um formato padrão que permita com que esses dados educacionais possam ser interpretados e armazenados em uma base de dados que dê suporte a um ambiente no qual várias funcionalidades poderão ser implementadas.

Para completar a camada responsável pelos dados na arquitetura Web-PIDE, o projeto utiliza a tecnologia de *Data Warehouse* (DW) e de *Data Marts* (DM) para sistematizar e armazenar os dados históricos com o objetivo de facilitar a tomada de decisão pelos gestores.

Essa tecnologia proporciona ao Ambiente de Apoio à Decisão (AAD), que no caso é a plataforma Web-PIDE, uma sólida e concisa integração dos dados para a realização de análises gerenciais, preocupando-se em integrar e consolidar as

informações de fontes heterogêneas e fontes externas, resumindo, filtrando e limpando esses dados, preparando-os para análise e suporte à decisão.

Com o uso das tecnologias propostas para construir a camada de dados, a plataforma Web-PIDE deve apresentar facilidades para:

- Extração de dados educacionais de fontes heterogêneas (existentes ou externas);
- Transformação e integração dos dados antes de seu armazenamento na plataforma Web-PIDE;
- Visualização dos dados em diferentes níveis. Os dados do DW podem ou não ser extraídos para um nível mais específico, que são os DMs e, a partir desses, para um banco de dados individual; e
- Utilização de ferramentas voltadas para acesso com diferentes níveis de apresentação.

Dada a facilidade que pode ser alcançada com a construção da plataforma Web-PIDE, acredita-se que ela será um elemento facilitador para a comunidade interessada em trabalhar com esses dados, bem como servir de incentivo para a condução e suporte de trabalhos acadêmicos, o que, atualmente, não possui um número representativo (PAIVA; TURINE; FORTES, 2006).

2.3.1 O formato dos dados disponibilizados para o projeto

Embora o objetivo inicial do projeto fosse tratar as bases de dados educacionais de todas as avaliações do INEP, usando os dados do ensino superior como ponto de partida, logo no início do projeto o INEP informou a necessidade de iniciar os estudos com as bases de dados do SAEB, pois é nesse contexto em que as análises são mais urgentes.

Assim, o INEP disponibilizou para o projeto Web-PIDE os microdados do SAEB em CD-ROM, sendo que as informações estão organizadas em:

- arquivos no formato ASCII (*American Standard Code for Information Interchange*);

- arquivos de entrada SAS (*Statistical Analysis Software*), que por sua vez definem as variáveis envolvidas nos microdados e são úteis por serem fáceis de manipular, possibilitando a leitura por outros software;
- documentação dos microdados, presente em arquivos nomeados “leia-me”, do tipo PDF (*Portable Document Format*).

As Figuras 2.8, 2.9 e 2.10 ilustram cada um dos formatos dos arquivos disponibilizados pelo INEP para o projeto.

```

DATA MATEMATICA_03ANO;
INFILE 'D:\DADOS\ALUNOS\MATEMATICA_03ANO.TXT' LRECL=971 MISSEVER;
INPUT
@1 MASCARA 8. /* MASCARA */
@9 ANO_MASCARA 8. /* ANO_MASCARA */
@17 ANO $char4. /* Ano de aplicação */
@21 SERIE $char2. /* série */
@23 DISC $char1. /* Disciplina */
@24 TURMA $char2. /* Número da turma */
@26 ALUNO $char3. /* Código do aluno */
@29 ESTRATO 8. /* Estrato da Amostra */
@37 UPA 12. /* Unidade Primária de Amostragem */
@49 DEP_ADM 8. /* Dependência
Administrativa(Estadual/Municipal/Particular) */
@57 LOCAL 8. /* Localização(Urbano/Rural) */
@65 REDE 8. /* Rede de Ensino(Pública/Particular) */
@73 UF 8. /* Unidade da Federação */
@81 UFSUD 8. /* Unidade da Federação para SUDAAN */
@89 REGIAO 8. /* Região */
@97 TAM_MUNIC 8. /* Tamanho do Município
@105 REG_METROP 8. /* Localizado em Região Metropolitana? */
@113 TAM_LCD 8. /* Tamanho/perfil da Cidade */
@121 TUR_BE 8. /* Turma tem aluno Bolsa Escola? */
@129 ALU_BE 8. /* Aluno tem Bolsa Escola? */
@137 PESO_AC 8.4 /* Peso calibrado(usado para expansão) */
@145 CADERNO 8. /* Caderno de Provas */
@153 BLOCO1 8. /* Bloco 1 do caderno */
@161 BLOCO2 8. /* Bloco 2 do caderno */
@169 BLOCO3 8. /* Bloco 3 do caderno */
@177 RESP_BL1 $char13. /* Respostas do aluno no Bloco 1 do Caderno */
@190 RESP_BL2 $char13. /* Respostas do aluno no Bloco 2 do Caderno */
@203 RESP_BL3 $char13. /* Respostas do aluno no Bloco 3 do Caderno */
@216 GAB_BL1 $char13. /* Gabarito do Bloco 1 do Caderno */
@229 GAB_BL2 $char13. /* Gabarito do Bloco 2 do Caderno */
@242 GAB_BL3 $char13. /* Gabarito do Bloco 3 do Caderno */
@255 PROFIC 12.5 /* Proficiência */
@267 ESTAGIO 8. /* Estágios de desempenho */
@275 A111_001 8. /* Sexo */
@283 A111_002 8. /* Como você se considera? */
@291 A111_003 8. /* Qual é o mês do seu aniversário? */
@299 A111_004 8. /* Qual é o ano do seu nascimento? */
@307 A111_005 8. /* Na sua casa tem televisão em cores? */
@315 A111_006 8. /* Na sua casa tem rádio? */
    
```

Figura 2.8. Início de um dos arquivos de entrada SAS do SAEB 2003.

22908909	2002200311H	2	1	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	3	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	4	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	7	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	9	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	11	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	13	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	15	22111	22908909	3	1	2	22	9	2	2	2
22908909	2002200311H	2	17	22111	22908909	3	1	2	22	9	2	2	2
22924234	2002200311H	1	1	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	3	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	5	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	7	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	9	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	11	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	13	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	15	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	17	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	19	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	21	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	23	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	25	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	27	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	29	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	31	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	33	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	35	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	37	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	39	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	41	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	43	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	45	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	47	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	49	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	51	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	53	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	55	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	57	15711	22924234	3	1	2	15	5	1	2	2
22924234	2002200311H	1	59	15711	22924234	3	1	2	15	5	1	2	2
22989307	2002200311H	1	1	51512	22989307	3	1	2	51	25	5	2	2

Figura 2.9. Estrutura de um dos arquivos ASCII do SAEB 2003.

Também foram disponibilizados para projeto os dados brutos do SAEB, que se diferenciam dos microdados apenas por conter informações de identificação do aluno e escola, sendo que esses dados são codificados e armazenados no atributo *máscara* dos microdados. Como essas informações não são relevantes para a criação da plataforma, o uso dos dados brutos foi descartado.

Nome da Variável	Descrição	Variáveis Categóricas				Início	Tamanho	Tipo
		Categoria	Descrição	Início	Tamanho			
		9	Estado					
		10	Região					
		11	Rio Grande do Norte					
		12	Paraná					
		13	Pernambuco					
		14	Alagoas					
		15	Sergipe					
		16	Bahia					
		17	Minas Gerais					
		18	Espírito Santo					
		19	Rio de Janeiro					
		20	São Paulo					
		21	Paraná					
		22	Santa Catarina					
		23	Rio Grande do Sul					
		24	Mato Grosso do Sul					
		25	Mato Grosso					
		26	Goiás					
		27	Distrito Federal					
REGIAO	Região	1	Norte		93	6	Numérica	
		2	Nordeste					
		3	Sudeste					
		4	Sul					
		5	Centro-Oeste					
TAM_MUNIC	Tamanho do Município	1	Munic. < 200.000 hab.		101	8	Numérica	
		2	Munic. >= 200.000 hab.					
REG_METROP	Localizado em Região Metropolitana?	1	Região Metropolitana		109	8	Numérica	
TAM_CID	Tamanho/perfil da Cidade	1	Cidade com pop. grande		117	8	Numérica	
		2	Cidade com pop. pequena					
TUR_BE	Turma tem aluno Bolsa Escola?	1	Turma TEM aluno BE		125	8	Numérica	
		2	Turma NAO tem aluno BE					
ALL_BE	Aluno tem Bolsa Escola?	1	Aluno TEM BE		133	8	Numérica	
		2	Aluno NAO tem BE					
PESO_AC	Peso calibrado(usado para expansão)	141			141	8	Numérica	
CADERNO	Caderno de Provas	149			149	8	Numérica	
BLOCO1	Bloco 1 do caderno	157			157	8	Numérica	
BLOCO2	Bloco 2 do caderno	166			166	8	Numérica	
BLOCO3	Bloco 3 do caderno	173			173	8	Numérica	
RESP_BL1	Respostas do aluno no Bloco 1 do Caderno	181			181	13	Alfanumérica	
RESP_BL2	Respostas do aluno no Bloco 2 do Caderno	194			194	13	Alfanumérica	
RESP_BL3	Respostas do aluno no Bloco 3 do Caderno	207			207	13	Alfanumérica	
GAB_BL1	Gabarito do Bloco 1 do Caderno	220			220	13	Alfanumérica	
GAB_BL2	Gabarito do Bloco 2 do Caderno	233			233	13	Alfanumérica	
GAB_BL3	Gabarito do Bloco 3 do Caderno	246			246	13	Alfanumérica	
PROFIC	Proficiência	259			259	12	Numérica	
ESTAGIO	Estágios de desempenho	271			271	15	Numérica	
		1	Muito Crítico					
		2	Crítico					
		3	Intermediário					
		4	Adequado					
		5	Avançado					
QUESTIONARIO 1								
A041_001	Sexo	1	Masculino		268	8	Numérica	
		2	Feminino					
A041_002	Como você se considera?	1	Branco(a)		294	8	Numérica	
		2	Pardo(a)					
		3	Pret(a)					
		4	Amarelo(a)					
		5	Indígena					
A041_003	Qual a sua idade?	1	8 anos ou menos		302	8	Numérica	

Figura 2.10. Estrutura do arquivo “Leia-me.pdf” do SAEB 2003.

2.4 Considerações finais

O INEP, desde 1991, aplica avaliações e censos nas instituições de ensino do país, com o intuito de retratar a realidade do setor educacional brasileiro.

Os dados oriundos desses diversos questionários são valiosas fontes para a definição de políticas de melhoria na educação e também para pesquisas em diversas áreas do conhecimento. No entanto, devido a diversos fatores, esses dados encontram-se distribuídos em diferentes bases de dados e de maneira não uniformizada, dificultando sua análise e reuso.

Apresentou-se neste capítulo o problema enfrentado atualmente pelo INEP, assim como suas avaliações, censos e pesquisas e a maneira como os resultados são apresentados e disponibilizados.

Foi descrito o projeto Web-PIDE, que ainda está em andamento e objetiva criar uma plataforma onde os interessados em conhecer e analisar os dados obtidos pelo INEP possam acessá-los de maneira organizada e de fácil manipulação. Um dos pilares para a construção dessa plataforma é a LIDE, que é apresentada no Capítulo 5.

Capítulo 3

VISUALIZAÇÃO DE INFORMAÇÕES

Este capítulo apresenta conceitos sobre Visualização de Informações, ilustra algumas técnicas e exemplifica com algumas ferramentas, o quão útil é a utilização de técnicas de visualização de informações para auxiliar o usuário a executar um vasto número de tarefas. Algumas pesquisas pertinentes ao assunto também são apresentadas.

3.1 Considerações Iniciais

A quantidade de informações armazenadas em repositórios computacionais de dados tem crescido mais rápido que a capacidade humana em explorá-los. Além disso, a complexidade de muitos conjuntos de dados supera a habilidade de identificar os conceitos subjacentes às informações armazenadas (MARQUES; MENDONÇA; SANTOS, 2004).

A área de Visualização de Informações, pioneiramente denominada por Tukey (1977) de Análise Exploratória Visual, é um campo de pesquisa dedicado à aplicação de técnicas de computação gráfica, comumente interativas, visando auxiliar o processo de análise e compreensão de um conjunto de dados, através de representações gráficas.

Assim, a área de Visualização de Informações se apresenta como um campo de estudo de grande utilidade, agregando técnicas que facilitem o entendimento de informações a partir de representações visuais (NASCIMENTO; FERREIRA, 2005).

Neste capítulo são descritos breves conceitos sobre Visualização de Informações na Seção 3.2 e técnicas de visualização, assim como exemplos, na Seção 3.3. Na Seção 3.4, são apresentadas algumas ferramentas e na Seção 3.5,

alguns trabalhos que envolvem visualização de ontologias são apresentados, por serem estes, dois temas correlatos a esta pesquisa. Na Seção 3.6 apresentam-se as considerações finais do capítulo.

3.2 Visualização de Informações

De acordo com Gershon, Eick e Card (1998), a Visualização é um processo de transformar os dados, informações e conhecimentos em forma visual, fazendo uso da capacidade visual natural dos seres humanos, fornecendo uma interface entre dois poderosos sistemas de tratamento da informação: a mente humana e o computador.

Com interfaces visualmente eficazes, é possível interagir rapidamente com grandes volumes de dados e descobrir características, padrões e tendências camufladas.

Tratando-se de análise de informações, a capacidade visual humana apresenta uma série de vantagens. Uma das principais vantagens é que uma grande quantidade de dados pode ser condensada em uma simples visualização, reforçando o famoso ditado popular “uma imagem vale mais que mil palavras”.

Como é citado por Nascimento e Ferreira (2005), outra vantagem é que o processo de Visualização envolve o sentido humano que possui maior capacidade de captação de informações por unidade de tempo. O sentido da visão é rápido e paralelo, sendo possível, inclusive, prestar atenção em um objeto de interesse especial, sem perder de vista o que está acontecendo ao redor.

Além disso, o sistema visual humano é treinado para reconhecer padrões, como se pode observar na Figura 3.1, em que a tarefa de identificar uma forma que está fora do padrão é fácil e rápida. Essas características permitem identificar e lidar com situações complexas que envolvem processamento visual.

As visualizações por si só trazem benefícios, uma vez que podem funcionar como uma extensão da memória humana e como um auxílio para o processo cognitivo. As imagens nos ajudam a entender o problema e/ou a encontrar uma solução para o mesmo.

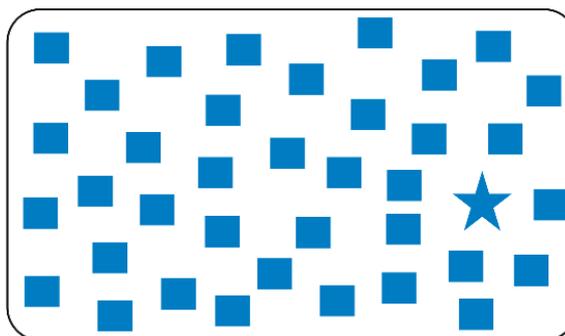


Figura 3.1. Reconhecimento de Padrões (NASCIMENTO; FERREIRA, 2005).

É possível observar no conjunto de Figuras 3.2(a), 3.2(b) e 3.2(c), que o mesmo conjunto de informações (gastos mensais de uma família) é apresentado de três maneiras distintas e, em cada forma de visualizar os dados, é possível observar e abstrair informações diferentes.

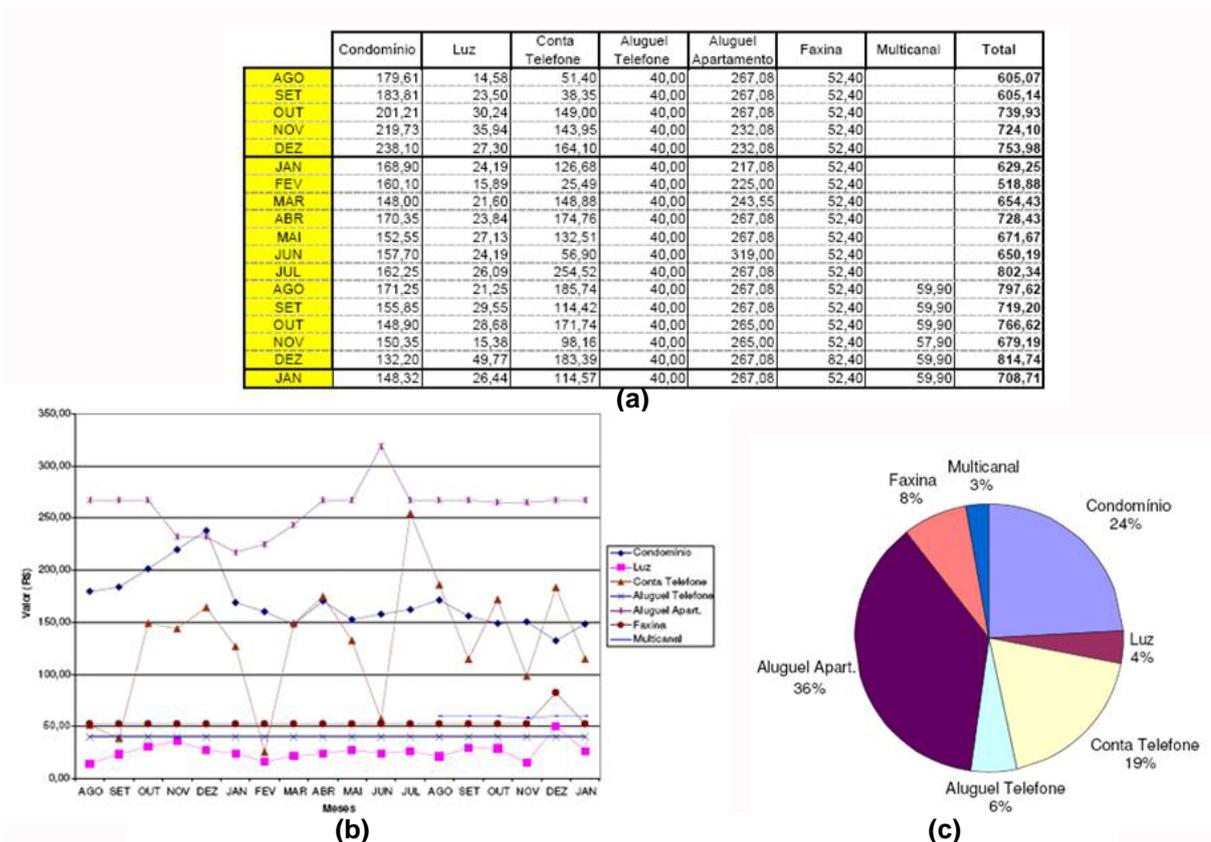


Figura 3.2. (a) Gastos apresentados em tabela. (b) Gráfico de linhas. (c) Gráfico de pizza (NASCIMENTO; FERREIRA, 2005).

A constatação de padrões ou de características visuais presentes em imagens contribuem de forma significativa para o processo de compreensão mais do que a simples observação dos dados em sua forma bruta.

Construir uma visualização que possibilite essa observação pode ser feita organizando os dados de acordo com algum critério e apresentando-os de modo visual. As visualizações possibilitam a recuperação de informações relevantes e a construção de novos conhecimentos.

É válido ressaltar a diferença entre Visualização de Informações e uma área similar chamada Visualização Científica. Ambas são destinadas ao processo de descoberta, percepção e tomada de decisões, com o objetivo de ampliar nossa atividade cognitiva. O objetivo é prover melhorias ao entendimento e aproveitamento do que é exposto, levando à aquisição e solidificação do conhecimento.

Na Visualização Científica os dados a serem apresentados geralmente correspondem a medidas de objetos físicos, fenômenos da natureza ou posições em um domínio espacial, possuindo, assim, uma representação geométrica, além de se centrar em um público de cientistas treinados. Como exemplo de visualizações nessa linha, pode-se citar a visualização de órgãos do corpo humano e de funções matemáticas.

Em Visualização de Informações, os dados são abstratos, como relacionamentos ou informações inferidas a partir da mensuração dos mesmos. Nesse caso, não há necessariamente uma representação geométrica inerente aos mesmos. A imagem deve ser gerada com base nos relacionamentos ou informações que podem ser inferidos acerca dos dados. A procura por uma visualização eficaz e de alto nível pode ser feita por usuários com níveis de conhecimento, interesses e necessidades diferentes (GERSHON; EICK; CARD, 1998).

A Figura 3.3 ilustra, de maneira simplificada, um processo automatizado de Visualização de Informações, em que uma imagem é produzida a partir de dados de entrada. Vale destacar que esse processo é chamado de Visualização de Informações e não Visualização de Dados, pois o interesse final é visualizar informações acerca da relação entre os dados, e não dos valores isolados.

Segundo Nascimento e Ferreira (2005), os dados abstratos a serem visualizados podem ser classificados em:

- Nominal: conjunto de elementos distintos, sem uma relação de ordem entre eles. Exemplo: João, Maria, Alberto;

- Ordinal: conjunto de elementos distintos, mas com uma relação de ordem entre os mesmos. Exemplo: Segunda, Terça, Quarta... , ou Janeiro, Fevereiro... ; e
- Quantitativo: faixa de valores numéricos. Essa categoria pode ser dividida em Intervalos (com valores discretos) e Razão (representando uma faixa contínua de valores).

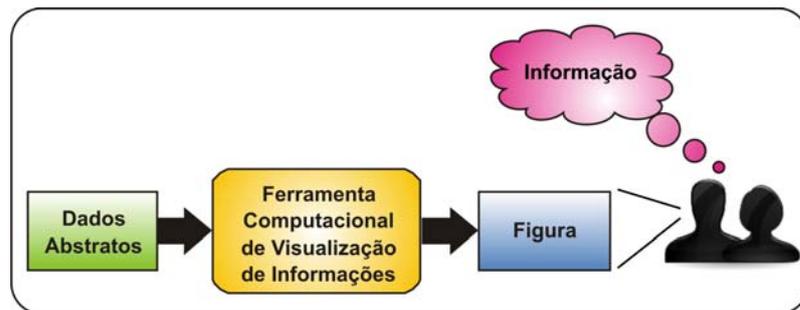


Figura 3.3. Processo simplificado de Visualização de Informações auxiliada por computador (NASCIMENTO; FERREIRA, 2005).

Essa classificação contribui na escolha de uma representação gráfica adequada para os dados.

A Figura 3.4 apresenta um modelo de referência para construir Visualização de Informações apresentado por Card e outros (Card *et al.*, 1999 *apud* Nascimento, Ferreira, 2005). Esse modelo prevê a divisão do processo de gerar uma imagem para um conjunto de dados em três etapas, sendo um refinamento do esquema da Figura 3.3 apresentada anteriormente.

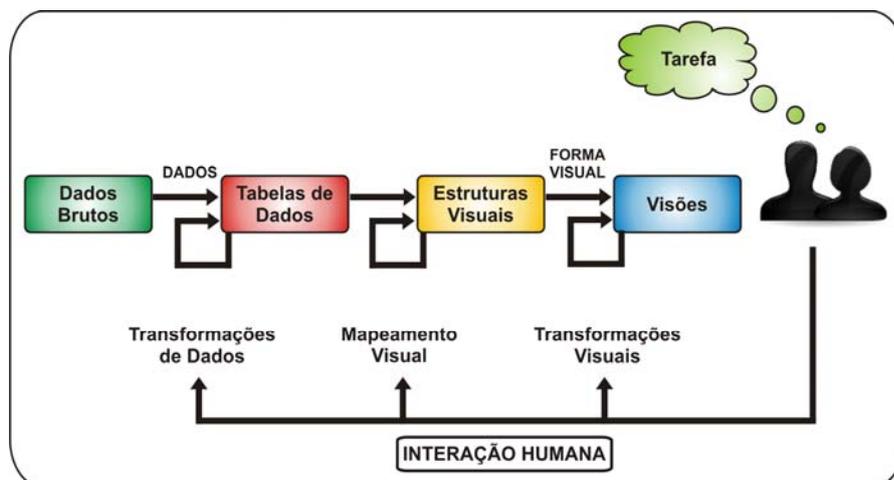


Figura 3.4. Modelo de Referência para Visualização (CARD *et al.*, 1999 *apud* NASCIMENTO; FERREIRA, 2005).

A primeira etapa é chamada de Transformações dos Dados. Nessa etapa, um conjunto de dados brutos é processado e organizado em uma representação lógica mais estruturada, geralmente na forma de uma ou mais tabelas em um banco de dados relacional. O processamento pode envolver a eliminação de dados redundantes, errados ou incompletos, bem como a filtragem e o agrupamento dos dados relevantes. Além disso, pode ser feita a inclusão de novas informações, como, por exemplo, de resultados de análises estatísticas (média, soma total, desvio padrão, etc.) realizadas sobre os dados brutos (NASCIMENTO; FERREIRA, 2005).

A segunda etapa é o Mapeamento Visual. Nessa etapa, o objetivo é construir uma estrutura visual que represente os dados da tabela. Toda estrutura visual pode ser decomposta em três partes: substrato visual, marcas e propriedades gráficas das marcas.

- Substrato visual: caracteriza o espaço para a visualização, sendo normalmente representado por eixos, tais como os eixos X e Y do plano cartesiano;
- Marcas visuais: são símbolos gráficos utilizados para representar os itens de dados, como figuras geométricas, linhas, áreas e volumes, como apresenta a Figura 3.5;
- Propriedades gráficas das marcas: são os atributos visuais que as caracterizam como ilustra a Figura 3.6.

Dessa maneira, o Mapeamento Visual consiste em associar os itens de dados a marcas visuais em um substrato visual. Cada atributo dos dados pode ser associado a propriedades gráficas das marcas.

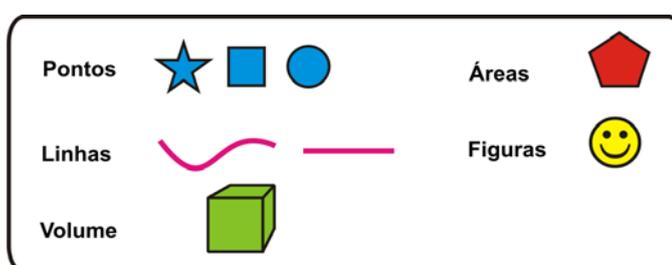


Figura 3.5 – Exemplo de marcas visuais. Adaptado de Nascimento e Ferreira (2005).



Figura 3.6 - Propriedades gráficas das marcas. Adaptado de Nascimento e Ferreira (2005).

A última etapa é a de Transformações Visuais, na qual é possível modificar e estender as estruturas visuais interativamente através de operações básicas como:

- Testes de localização, que possibilitam obter informações adicionais sobre um item da tabela de dados;
- Controles de ponto de vista, os quais permitem ampliar, reduzir e deslocar a imagem com o objetivo de oferecer visões diferentes; e
- Distorções da imagem, visando criar ampliações de uma região específica em detrimento de outra.

Uma característica importante nessa etapa é que os mecanismos de interação implementados permitem ao usuário explorar diferentes cenários para um melhor entendimento dos dados visualizados.

O esforço de exploração dos dados é repassado parcialmente para o computador, uma vez que os cálculos e o redesenho da imagem são realizados pela máquina. Ao usuário, cabe a tarefa de observar o que acontece quando a Visualização é modificada (NASCIMENTO; FERREIRA, 2005).

Como mencionado por Gershon, Eick e Card (1998), a importância da Visualização de Informações e seu potencial para resolver problemas do mundo real está longe de ser um interesse meramente acadêmico.

Como exemplo de aplicações comerciais que envolvem a disciplina de Visualização de Informações, pode-se citar a resolução de problemas no mercado financeiro (WRIGHT, 1997) e também o uso, pela Xerox PARC, de idéias desenvolvidas em laboratório (*Perspective Wall, Cone Tree, Wide Widgets*) para traçar uma nova geração de interfaces melhoradas (RAO, 1997).

Um exemplo pertinente a esta pesquisa é dado por Becker (1997), que utiliza Visualização de Informações para representar conhecimentos extraídos de grandes

bases de dados, usando mineração de dados para apoiar a tomada de decisão em ambientes comerciais e outros.

3.3 Técnicas de Visualização

As técnicas de visualização são baseadas numa representação visual e em mecanismos de interação que possibilitam ao usuário manipular essa representação, de modo a melhorar a compreensão do conjunto de dados representado. O nível de abstração dessa representação é mais alto porque, frequentemente, não há relação direta entre os dados e uma entidade física ou geométrica, ou o interesse do usuário não está em dados brutos, mas em observar características ou padrões no conjunto (FREITAS; CHUBACHI; LUZZARDI, 2001).

Considerando o objetivo das técnicas de visualização, pode-se considerá-las, essencialmente, um processo de mapeamento. Esse mapeamento se dá entre as representações computacionais e as representações perceptíveis. Assim, esse mapeamento deve ser feito de forma a maximizar a comunicação e o entendimento humano (CEMIN, 2001).

De acordo com Luzzardi (2003), as técnicas de visualização de informações utilizam representações ou metáforas visuais para exibir graficamente dados que geralmente não possuem representação direta, óbvia e natural.

Uma representação visual é uma figura ou imagem que representa graficamente um conjunto de dados e auxilia o usuário a perceber visualmente características desses dados. Uma metáfora visual utiliza uma representação visual e, na maioria das vezes, mecanismos de interação, que incrementam o processo de percepção por adaptar os dados conforme acontece o processo cognitivo.

Nas várias técnicas, frequentemente, os autores buscaram inspiração em objetos do mundo real (ou geométricos), para mapear o conjunto de informações. Uma técnica de visualização envolve:

- Construção de um modelo dos dados;
- Seleção de algum mecanismo de mapeamento do modelo em um objeto de visualização abstrato, como uma imagem ou mapa de contornos;

- Geração da imagem no dispositivo gráfico de exibição.

De acordo com Freitas e Wagner (1993), é possível criar inúmeros modelos, porém, deve-se escolher o mais adequado, uma vez que este afeta diretamente a interface da visualização. É importante frisar que, de acordo com o tipo dos dados e da tarefa a ser realizada pelos usuários, se pode identificar qual é a representação visual mais adequada.

A seguir, são descritas e exemplificadas algumas técnicas de visualização comumente utilizadas.

3.3.1 Coordenadas Paralelas

A técnica *Coordenadas Paralelas* foi proposta por Inselberg e Dimsdale (1990). Essa técnica associa as dimensões dos dados a eixos paralelos verticais equidistantes, denominados coordenadas, possibilitando a criação de representações gráficas multidimensionais.

Os eixos verticais representam as dimensões ou atributos dos dados. Uma linha representando cada item dos dados conecta os eixos nos seus respectivos valores, o que permite a observação de padrões e tendências. A Figura 3.7 ilustra um dos vários exemplos de emprego dessa técnica, apresentados por Inselberg e Dimsdale (1990).

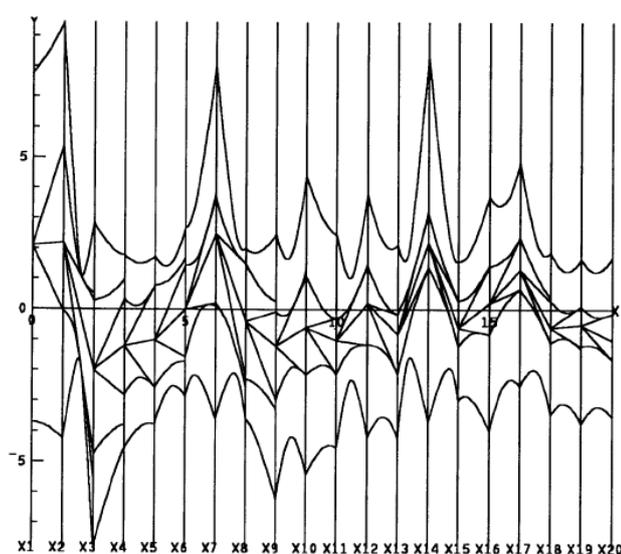


Figura 3.7. Exemplo da técnica Coordenadas Paralelas por Inselberg e Dimsdale (1990).

Outro exemplo da aplicação dessa técnica pode ser analisado em Pousman e Stasko (2006), como ilustra a Figura 3.8. Os autores usam a técnica para visualizar dados de dezenove sistemas de informação, traçando linhas para cada sistema e classificando-os ao longo da concepção das dimensões. Cores diferentes são utilizadas para designar grupos de sistemas que são semelhantes, possibilitando com isso, abstrair um novo tipo de informação sobre os sistemas.

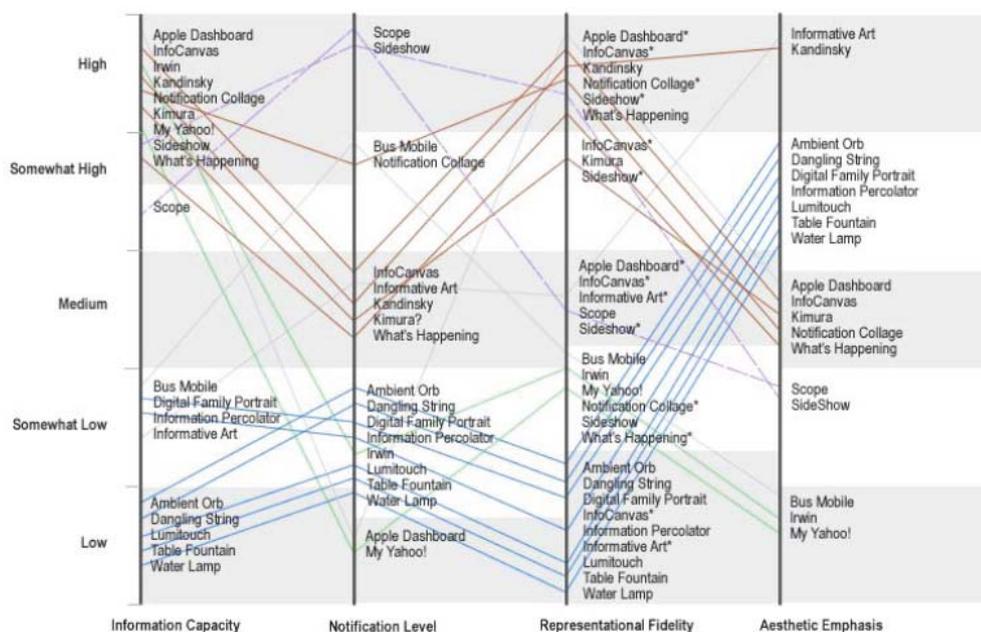


Figura 3.8. Exemplo de utilização da técnica Coordenadas Paralelas (POUSMAN, STASKO, 2006).

3.3.2 Glifos

A técnica de Glifos (*Glyphs*), também chamada de Ícones, assim como as Coordenadas Paralelas, é utilizada para a visualização de dados multidimensionais e podem ser compostos por atributos geométricos, tais como forma, tamanho, orientação, posição ou direção, e atributos de aparência, como cor, textura e transparência, todos determinados pelos valores dos atributos da entidade.

Em geral, características e valores diversos são exibidos em gráficos ou mapas dos mais variados tipos, desde os tradicionais gráficos de pontos ou linhas e mapas utilizando cores, a conjuntos de ícones, ou glifos, dispostos de acordo com o espaço-domínio. Ícones, em geral, são utilizados para identificar uma entidade ou elemento amostrado num contexto, representando, na maioria das vezes, a característica principal dessa entidade ou amostra (FREITAS; CHUBACHI; LUZZARDI, 2001).

Chernoff (1973) é um dos pioneiros dessa técnica. O autor observou que os seres humanos são sensíveis a uma grande variedade de expressões faciais. Ele, então, sugeriu a utilização de ícones representando faces, sendo que algumas características como o tamanho dos olhos, a altura da sobrancelha e a forma da boca pudessem ser associados a cada atributo dos dados. Para exemplificação, aplicou sua técnica para estudar exemplos geológicos. Sua técnica, conhecida como Faces de Chernoff, pode ser observada na Figura 3.9. Essa é uma forma que permite o usuário compreender rapidamente as informações relevantes do conjunto de dados. Muitos autores investigam o uso dessa técnica e de suas diversas variações e aprimoramentos. Um exemplo de utilização da técnica de Glifos pode ser observado na Figura 3.10, em que a ferramenta PDBVis utiliza a técnica para disponibilizar a visualização de um banco de dados de proteínas.

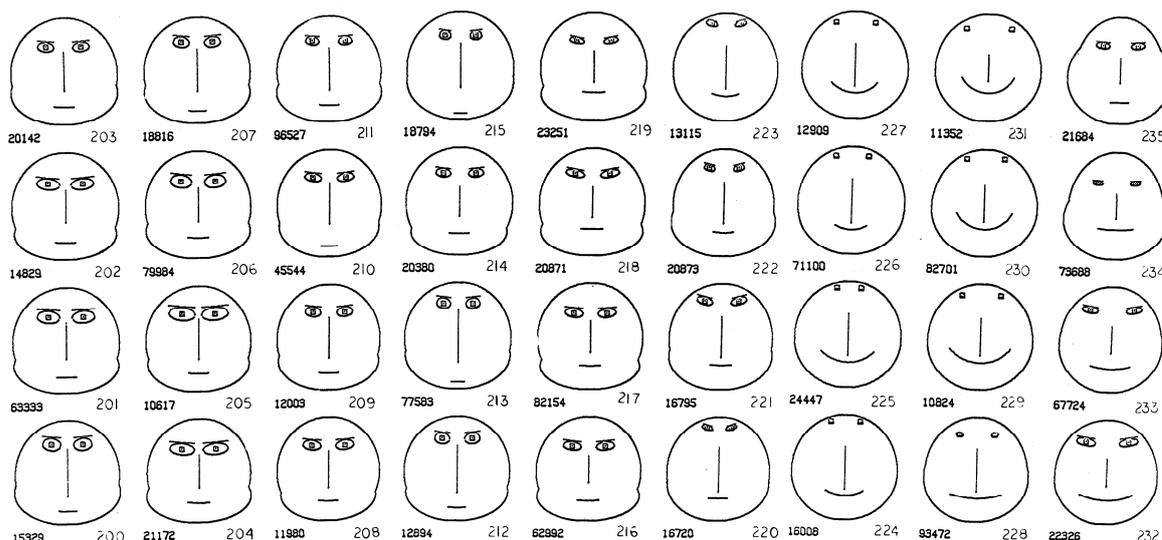


Figura 3.9. Faces de Chernoff (CHERNOFF, 1973).

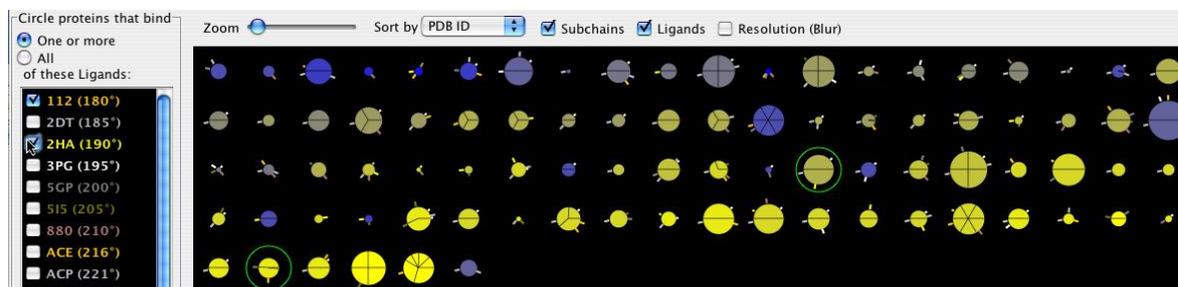


Figura 3.10. Tela da ferramenta PDBVis (Karve, Gleicher, 2007).

3.3.3 Grafos

Os grafos são estruturas abstratas, utilizadas para modelar informações, podendo ser usados para representar qualquer informação que possa ser modelada como objetos e as ligações entre os mesmos. Por isso, quase todos os domínios utilizam grafos de alguma forma para representar dados (TOLLIS, 1996).

Tollis (1996) afirma que na área de Visualização de Informações, a utilidade de um desenho de grafos depende da sua legibilidade, isto é, a capacidade de transmitir o significado do diagrama de forma rápida e clara.

O desenvolvimento de algoritmos para manipular e desenhar grafos é amplamente investigado por pesquisadores, principalmente por ser flexível e permitir sua utilização em diversas de situações. Várias normas gráficas têm sido propostas. Normalmente, os vértices são representados por círculos ou caixas, e as arestas são representadas por uma curva entre os vértices.

Na Figura 3.11 é possível observar a visualização, por meio de um simples grafo, de rede de telecomunicações. Na Figura 3.12 é possível observar um dos vários exemplos encontrados na internet da utilização dessa técnica. Nesse caso, trata-se de um site sobre informações de música e cinema, nomeado de Live Plasma (LIVEPLASMA, 2009).

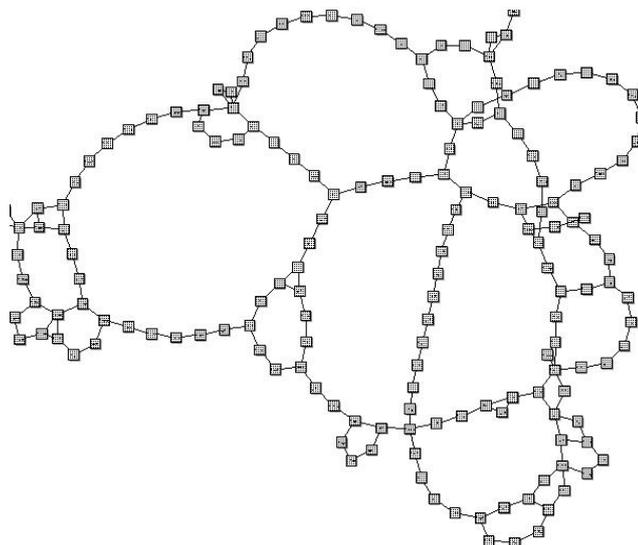


Figura 3.11. Exemplo de utilização da técnica de Grafos (TOLLIS, 1996).



Figura 3.12. Tela de um site que disponibiliza o resultado da busca do usuário por meio de grafos (LIVEPLASMA, 2009).

3.3.4 Árvore Hiperbólica

O conceito central das técnicas que utilizam a abordagem de interação Foco+Contexto é apresentar uma visão geral dos dados a serem visualizados, mas destacando uma região de interesse (foco) através de uma ampliação suave da mesma. Geralmente, procura-se ampliar a região de interesse, enquanto se compacta o restante da imagem (do contexto).

A técnica de Árvore Hiperbólica (*Hyperbolic Tree*) combina Foco+Contexto com desenho radial de árvores para auxiliar na exploração de grandes hierarquias. Um dos principais trabalhos relacionados a essa técnica é a proposta de Lamping e Rao (1996), a qual é possível observar por meio da Figura 3.13.

Uma propriedade conveniente desse plano é que a circunferência de um círculo, cujo centro coincide com o centro da tela, cresce exponencialmente com o seu raio. Com isso, há mais espaço disponível para se compactar informações à medida que se aproxima da periferia do desenho.

Assim como acontece com outras técnicas, a evolução e aprimoramento da visualização é constante. Percebe-se que quanto mais utilizada, mais variações e melhorias são acopladas às técnicas. A Figura 3.14 ilustra uma comum utilização da técnica em questão, o *Browser Hiperbólico*.

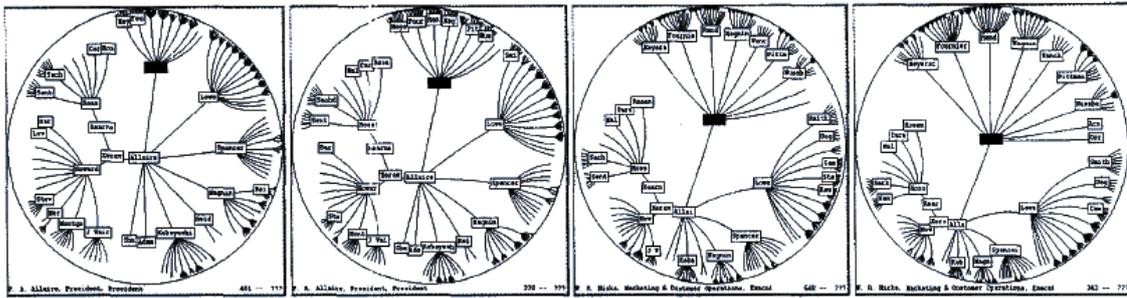


Figura 3.13. Proposta de Árvore Hiperbólica de Lamping e Rao (Lamping, Rao, 1996).

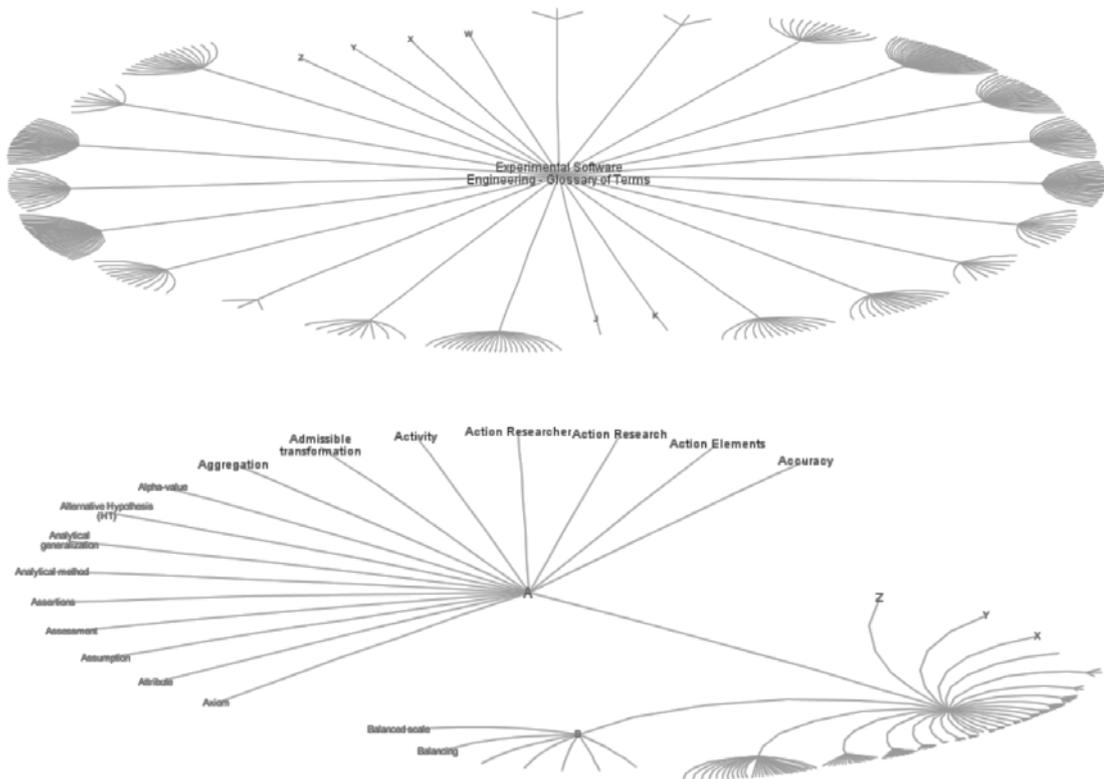


Figura 3.14. (a) *Browser* hiperbólico utilizado para visualizar um glossário de ESE. (b) Visualização dos termos que iniciam com a letra A no glossário de ESE (ESEGLOSSARY, 2009).

Nesse exemplo, a técnica é utilizada para disseminar entre pesquisadores e interessados na área de Engenharia de Software Experimental (ESE) um glossário, com termos e definições pertinentes ao assunto.

Cada aresta da visualização corresponde a uma letra do alfabeto. Ao clicar em cada aresta, é possível visualizar os termos que iniciam com essa letra e, ao clicar em cada termo, o usuário é remetido a uma página que contém sua definição. Na Figura 3.14(a) é ilustrado o glossário na sua visualização inicial e na Figura 3.14(b), a visualização após clicar na letra A.

3.3.5 Fish-Eye

Apresentada por Furnas (1982), a técnica de *Fish-Eye* (olho de peixe) também segue a abordagem de interação Foco+Contexto, assim como a Árvore Hiperbólica, descrita anteriormente.

Na ocasião, o autor incomodava-se com o fato dos usuários, muitas vezes, visualizarem grandes estruturas, como linhas de código estruturado, bases de dados, organogramas, linhas de texto, menus de acesso de sistemas e outros, através de telas pequenas, como vídeos de 24x80 caracteres. Para Furnas (1982), era necessário permitir a visualização dos detalhes, mas não perder de vista o conteúdo global. Na Figura 3.15 é ilustrado o exemplo oferecido pelo autor, no qual se tem a visualização do código-fonte de um programa na linguagem C.

Na Figura 3.15(a) é apresentada a visualização padrão. O autor ressalta que ao querer analisar a linha de código 39, informações importantes para o entendimento desta linha, como por exemplo, o bloco de instrução que inicia a estrutura *switch case*, não são visualizados. Por meio da técnica *Fish-Eye*, na Figura 3.15(b), é possível visualizar o contexto da instrução *for* da linha 39 (visualizando todas as condições da instrução *switch case*), sendo que para isso, foram omitidas as linhas de código das estruturas *for* que obedeciam as outras condições da instrução *switch case*.

<pre> 28 t[0] = (t[0] + 10000) 29 - x[0]; 30 for(i=1;i<k;i++){ 31 t[i] = (t[i] + 10000) 32 - x[i]; 33 (1 - t[i-1]/10000); 34 t[i-1] %= 10000; 35 } 36 t[k-1] %= 10000; 37 break; 38 case 'e': >39 for(i=0;i<k;i++) t[i] = x[i]; 40 break; 41 case 'q': 42 exit(0); 43 default: 44 noprint = 1; 45 break; 46 } 47 if(!noprint){ 48 for(i=k - 1;t[i] <= 0 && i > 0;i--); 49 printf("%d",t[i]); 50 if(i > 0) { </pre>	<pre> 1 #define DIG 40 2 #include <stdio.h> ...4 main() 5 { 6 int c, i, x[DIG/4], t[DIG/4], k = DIG/4, noprint = ...8 while((c=getchar()) != EOF){ 9 if(c >= '0' && c <= '9'){ ...16 } else { 17 switch(c){ 18 case '+': ...27 case '-': ...38 case 'e': >>39 for(i=0;i<k;i++) t[i] = x[i]; 40 break; 41 case 'q': ...43 default: ...46 } 47 if(!noprint){ ...57 } 58 } 59 noprint = 0; 60 } 61 } </pre>
--	---

(a)

(b)

Figura 3.15. (a) Visualização de código-fonte sem a técnica fish-eye. (b) Exemplificação de Furnas para técnica Fish-Eye (FURNAS, 1982).

É grande o número de situações em que a técnica *Fish-Eye* pode ser útil. Há uma grande quantidade de sites na internet que ensinam a implementar essa técnica, permitindo que muitas pessoas possam inseri-la em suas páginas pessoais.

No site do Laboratório de Interação Humano-Computador da Universidade de Maryland (<http://www.cs.umd.edu/hcil/fisheyemenu/fisheyemenu-demo.shtml>) é possível ver uma versão de um menu *Fish-Eye*, a aplicação mais encontrada da técnica, que contém a lista de 266 países. O usuário pode percorrer essa lista, ou escolher outras opções que o menu oferece. Na Figura 3.16 é possível observar o menu mencionado.

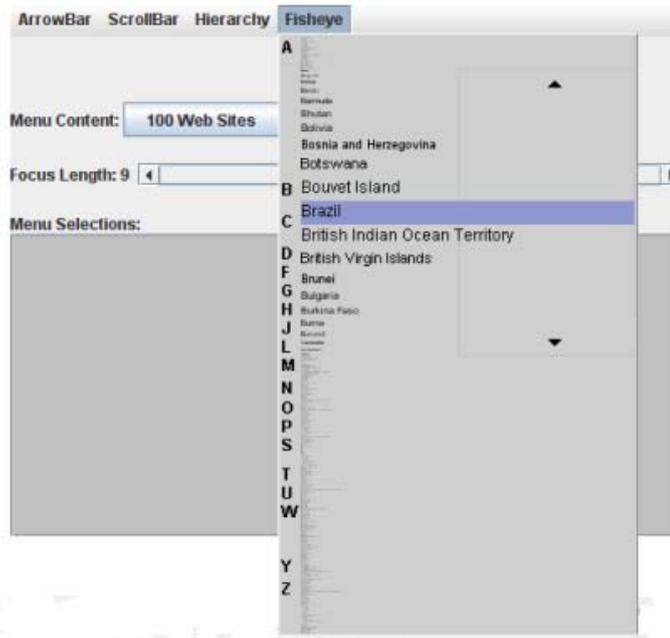


Figura 3.16. Menu *Fish-Eye* (FISHEYEMENU, 2009).

3.3.6 *Tree-Map*

Motivados pela falta de ferramentas para a visualização das grandes estruturas de diretório de unidades de disco rígido, Johnson e Shneiderman (1991) desenvolveram a técnica *Tree-Map*, que faz uso de 100% do espaço disponível na tela para mostrar uma estrutura hierárquica.

Na Figura 3.17 é possível observar a visualização gerada pelos autores para visualizar a estrutura de diretório de um disco rígido.

A técnica consiste em representar o nível mais alto da hierarquia como uma região retangular que preenche todo o espaço de desenho. Os níveis mais baixos são desenhados recursivamente como retângulos dentro da região maior. O tamanho de cada retângulo é proporcional ao número de itens nos níveis imediatamente abaixo na hierarquia.

De acordo com Johnson e Shneiderman (1991), uma grande quantidade de

informações são estruturadas hierarquicamente. A maioria das pessoas entende facilmente esta organização quando se trata de estruturas pequenas, no entanto, essa facilidade diminui na mesma proporção que as estruturas aumentam.

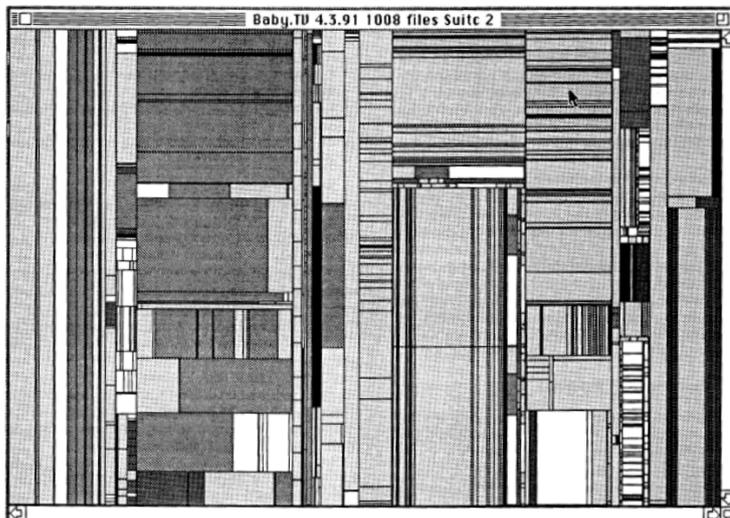


Figura 3.17. Visualização da estrutura de diretórios de um disco rígido (JOHNSON, SHNEIDERMAN, 1991).

Um exemplo dessa técnica pode ser visto em um site de notícias chamado Newsmap (NEWSMAP, 2009), que utiliza a *Tree-Map* para mostrar ao usuário notícias mundiais, permitindo que este interaja com a visualização, filtrando as notícias por tema, data e país. Na Figura 3.18 é possível observar a tela inicial do site mencionado.



Figura 3.18 – Site de notícias que utiliza a técnica *Tree-Map* (NEWSMAP, 2009).

3.4 Ferramentas de Visualização

Por oferecer grandes vantagens, várias ferramentas utilizam técnicas de visualização. Como exemplo, pode-se citar a ferramenta TreeMap, desenvolvido por Ben Shneiderman e sua equipe do Laboratório de Interação Homem-Computador da Universidade de Maryland. Como se pode observar na Figura 3.19, a ferramenta implementa a técnica *Tree-Map*.

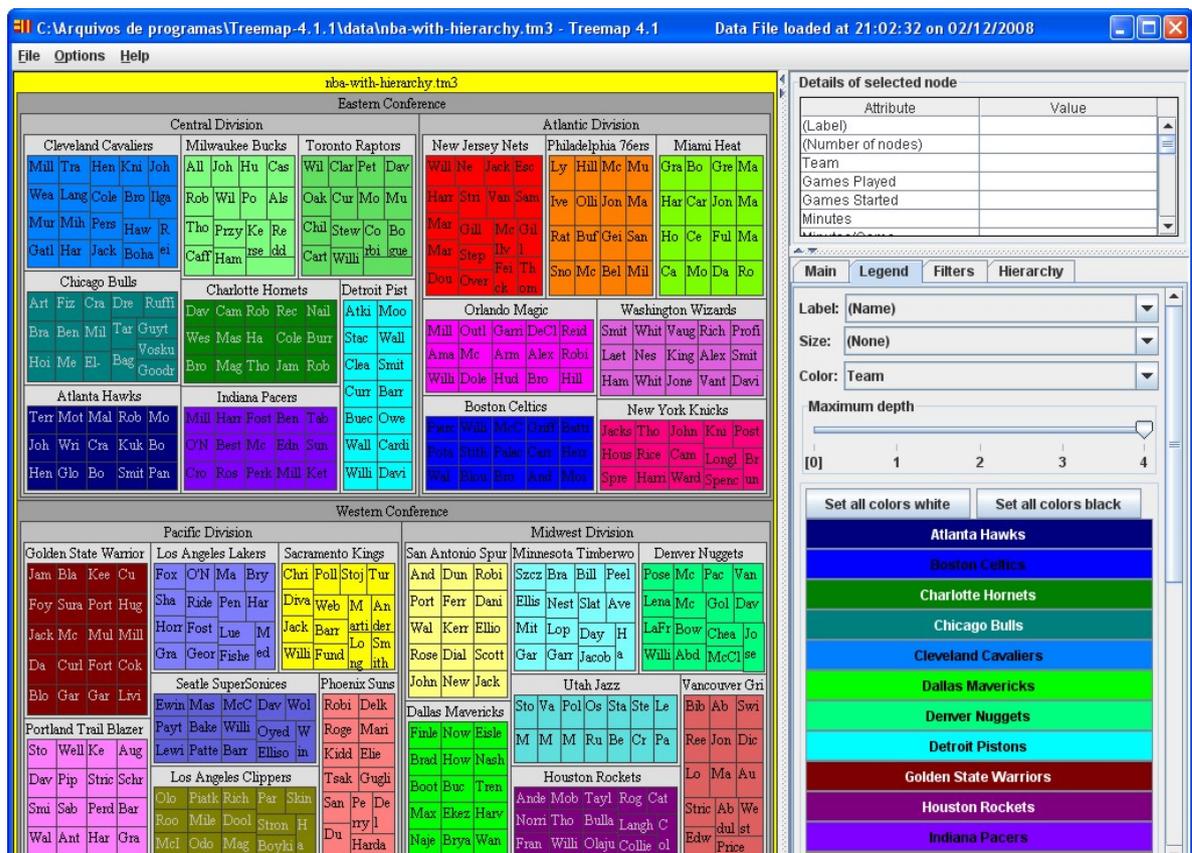


Figura 3.19. Ferramenta TreeMap.

Outra ferramenta que utiliza essa técnica é a CRISTA (*Code Reading Implemented with Stepwise Abstraction*) de Porto, Mendonça e Fabbri (2008), que auxilia na inspeção de código-fonte de programas. Na Figura 3.20 é exemplificada a tela da ferramenta.

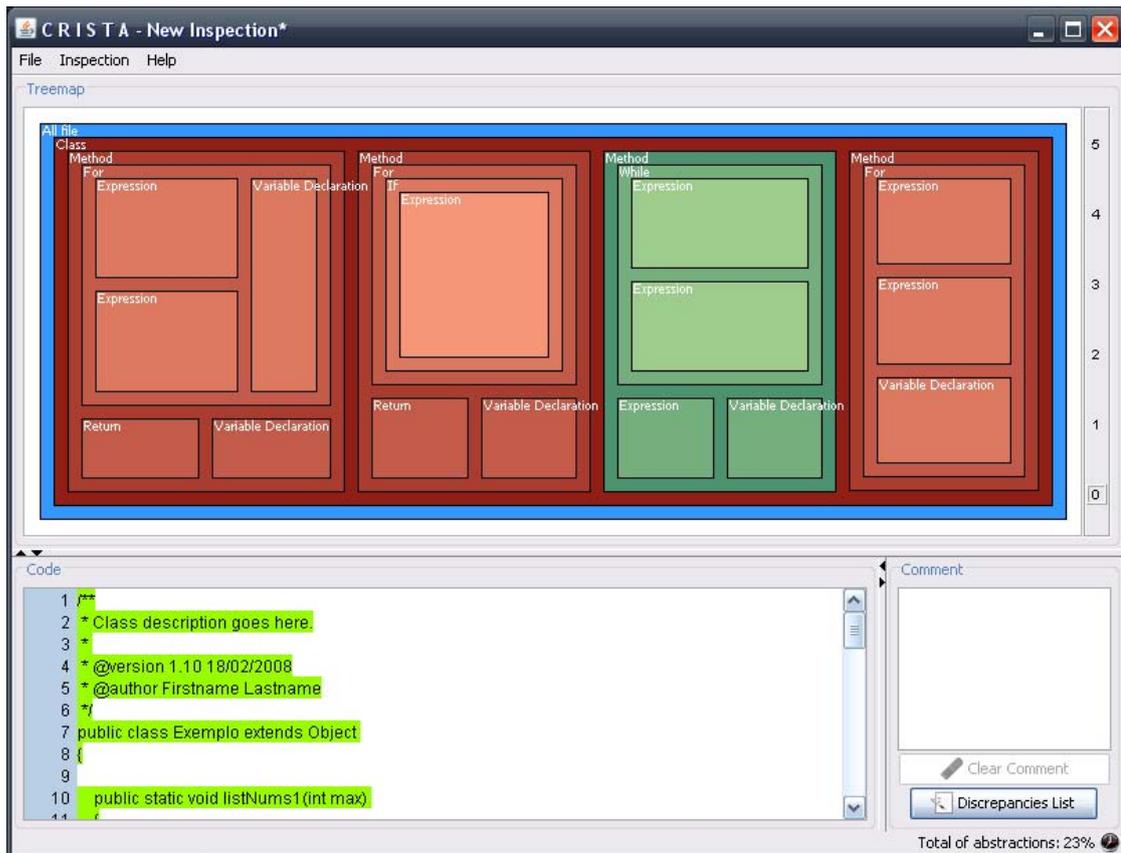


Figura 3.20. Tela da ferramenta CRISTA (PORTO; MENDONÇA; FABBRI, 2008).

A ferramenta JFLAP, exemplificada por meio da Figura 3.21, utiliza a técnica de grafos para disponibilizar a visualização dos autômatos que ela auxilia o usuário a criar e testar.

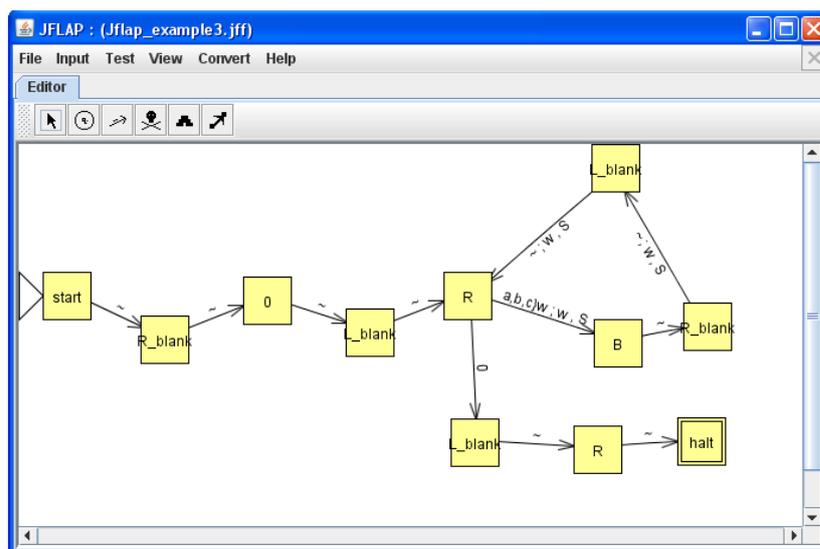


Figura 3.21. Tela da ferramenta JFlap (JFLAP, 2009).

Outra ferramenta que disponibiliza vários recursos de visualização é a Tulip (TULIP, 2009). Na Figura 3.22 é possível observar uma tela da ferramenta representando por meio de diversas técnicas de visualização, a estrutura de diretórios de um disco rígido.

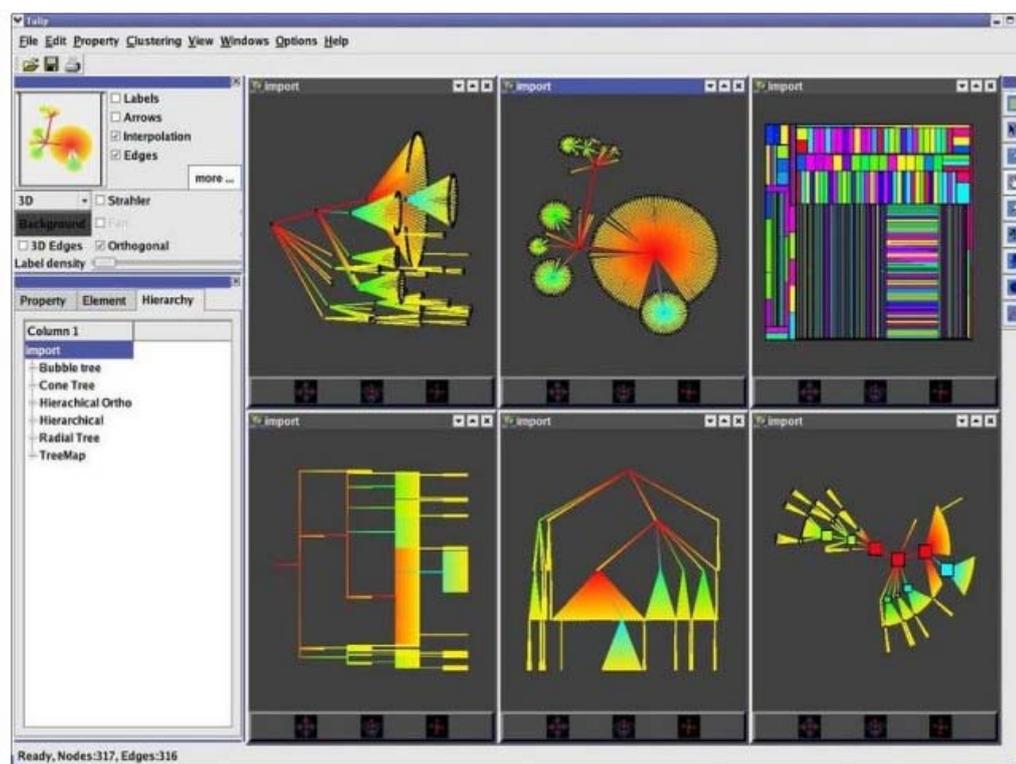


Figura 3.22. Tela da ferramenta Tulip (TULIP, 2009).

Atualmente, muitas bibliotecas e pacotes de software encontram-se disponíveis, permitindo que programadores de software e/ou usuários de computador criem suas próprias visualizações. Alguns exemplos são:

- Graphviz – programa de código aberto para desenho de grafos desenvolvido na AT&T (GRAPHVIZ, 2009);
- JGraphT – outra biblioteca livre de classes em Java para desenho de grafos (JGRAPHT, 2009);
- Hypergraph – pacote para visualização de grafos em um browser hiperbólico (HYPERGRAPH, 2009);
- XmdvTool – programa aberto para visualizar dados multidimensionais (XMDV, 2009);

- OpenDX - ambiente aberto para visualização científica desenvolvido pela IBM (OPENDX, 2009);
- Prefuse – biblioteca gratuita que permite aplicar diversas técnicas de visualização em aplicações desenvolvidas em Java (PREFUSE, 2009).

3.5 Trabalhos relacionados

Inúmeros trabalhos são propostos contextualizando o uso de Visualização de Informações em diversas áreas. Alguns exemplos puderam ser observados na seção anterior, na qual ferramentas e sites destinados a diferentes situações foram exemplificados.

Sendo o uso de visualização destinado, principalmente, para abstrair informações sobre um conjunto de dados, pesquisas que objetivam propor técnicas e ferramentas de visualização para adquirir conhecimento são realizadas e aprimoradas constantemente.

Nesse contexto, pode-se mencionar a pesquisa de Chen, Kuljis e Card (2001), que propõem a visualização de coleções de referências bibliográficas de trabalhos científicos para, a partir destes, descobrir domínios de conhecimento implícitos no conjunto de referências. Os autores utilizam a ferramenta PathFinder, que por sua vez utiliza PFNETs, que é um método estrutural de cognição desenvolvido por psicólogos na década de 80.

Auvil e outros (2007) apresentam um estudo em que utilizam quatro ferramentas de mineração de dados que possuem recursos de visualização: D2K (*Data To Knowledge*), RiverGlass e ReConTM, FeatureLens e DISCUS.

Os autores visualizam, nas quatro ferramentas, um conjunto de dados denominado VAST, composto de notícias, entradas de blogs, pequenos mapas e tabelas de dados. Em cada ferramenta são executadas tarefas que permitem abstrair novas informações e refinar o conjunto de dados.

Na D2K foram detectadas e removidas duplicações de dados. Na RiverGlass, aliada com ReConTM, os documentos resultantes da atividade anterior foram visualizados e analisados de acordo com a semelhança entre eles. A ReConTM também foi usada para extrair, automaticamente, entidades comuns nos

documentos, permitindo a identificação de possíveis relações entre pessoas e organizações. Uma característica importante da ferramenta ReConTM é a utilização de ontologias construídas para segmentar o terrorismo, crimes violentos e narcóticos.

A ferramenta FeatureLens foi utilizada para aplicar três padrões de análises nos dados resultantes do uso da ferramenta D2K. A FeatureLens permite aos usuários visualizar ocorrências duplicadas dos padrões aplicados. Uma cor diferente é atribuída a cada padrão, agregando valor à atividade de identificação de padrões realizada com esta ferramenta.

Para finalizar a pesquisa, os autores usaram a ferramenta DISCUS para auxiliar a sumarização dos conceitos-chave das relações e fatos abstraídos no conjunto de dados por meio das atividades anteriores.

Outra pesquisa que utiliza a visualização para abstrair informações de um conjunto de dados é a de Ichise, Satoh e Numao (2009). Com o intuito de identificar novos domínios de conhecimento interdisciplinar, os autores utilizaram uma ferramenta denominada Pajek para gerar a visualização dos dados em grafos. Foi utilizado um conjunto de dados sobre pesquisas e pesquisadores japoneses. Esses dados são submetidos a um órgão do governo para que as pesquisas recebam apoio financeiro. Como há um padrão para submissão, tem-se um grande conjunto de dados padronizados, que contém detalhes relevantes sobre o objetivo das pesquisas e sobre quem irá executá-las.

A visualização desses dados através de grafos permite visualizar, por meio dos relacionamentos gerados, pesquisas correlatas e pesquisas que se entrelaçam, evidenciando, assim, novas disciplinas a serem investigadas.

Como apresentado neste capítulo, o uso de metáforas visuais para visualizar grande quantidade de dados e informações é de grande utilidade, e muitas ferramentas foram desenvolvidas para auxiliar nessa tarefa.

Nesse contexto, as ontologias (tema do Capítulo 6) podem ser vistas como um grande conjunto de dados e o uso de visualização para observá-las também pode apresentar benefícios.

Katifori e outros (2007) apresentam uma investigação sobre diversas ferramentas desenvolvidas para visualizar ontologias, apresentando suas características e relacionando-as com um conjunto de exigências elaboradas para esse tipo de visualização. Destaca-se que a maioria dessas ferramentas são plug-

in's para ferramentas específicas de criação e edição de ontologias, que serão abordadas no Capítulo 4.

A Tabela 3.1 apresenta as ferramentas para visualização de ontologias mencionadas pelos autores, que as categorizam de acordo com as possibilidades de navegação e interação que eles oferecem. Nesta pesquisa foi utilizada a ferramenta Jambalaya, que é destacada na tabela.

Tabela 3.1. Ferramentas de Visualização de ontologias. Adaptada de Katifori e outros (2007).

Retração e expansão dos nós	Movimentação e/ou rotação do grafo	Movimentação e/ou rotação do ponto de visão	Zoom	Janela de visão geral	Opções de voltar e avançar	Animação nas transições
Class Browser	Class Browser	Tree Viewer	OntoViz	OntoViz	Jambalaya	SpaceTree
OntoViz	OntoViz	BiFocal Tree	IsaViz	Isa Viz	Information Pyramids	OntoTrack
SpaceTree	IsaViz	OntoSphere	SpaceTree	OntoTrack	CropCircles	Cone Tree
OntoTrack	SpaceTree	Beam Trees	OntoTrack	Jambalaya		fsviz
GOBar	OntoTrack	TGVizTab	Cone Tree	Information Pyramids		Reconfigurable Disk Tree
GOSurfer	Cone Tree	3D Hyperbolic Browser	fsviz	TreePlus		Jambalaya
Cone Tree	fsviz		Reconfigurable Disk Tree			TGVizTab
fsviz	Reconfigurable Disk Tree		Tree Viewer			OZONE
Reconfigurable Disk Tree	OntoSphere		OntoSphere			BiFocal Tree
Information Slices	Jambalaya		Grokker			3D Hyperbolic Browser
TGVizTab	Information Pyramids		Jambalaya			TreePlus
BiFocal Tree	TGVizTab		Information Cube			
TreePlus	Gopher VR		Information Pyramids			
	Harmony Information Landscape		CropCircles			
	fsn		TreeMap			
			Sequoia View			
			BeamTrees			
			TGVizTab			
			Gopher VR			
			Harmony Information Landscape			
			fsn			

Como pôde ser observado, a eficácia da visualização para amparar a identificação de novas informações e para tornar possível a visualização total de um grande volume de dados é comprovada por muitas pesquisas, além de ser um objeto de estudo constante.

3.6 Considerações Finais

Seja para analisar dados por meio de gráficos dos mais variados tipos, auxiliar na mineração e comparação de dados, visualizar linhas de código-fonte, visualizar resultado de consultas realizadas, exibir relacionamentos ou facilitar a exploração de um extenso conjunto de dados, o uso da Visualização de Informações pode ser aplicado em diversas áreas e proporcionar benefícios aos usuários, que

conseguem interagir de maneira mais fácil e observar características e/ou tendências camufladas rapidamente.

A escolha de uma técnica adequada para o tipo dos dados que serão analisados é de fundamental importância, assim como a escolha da ferramenta a ser utilizada, que deve proporcionar um ambiente eficiente e que possibilite interações do usuário.

Usar visualização para analisar um grande conjunto de dados permite uma visão ampla dos dados e possibilita a abstração de novas informações sobre o que está sendo visualizado de forma mais rápida, o que exigiria um esforço maior dos usuários se a análise fosse feita manualmente. Até mesmo em casos em que o conjunto de dados é menor, a visualização oferece vantagens, pois por meio de uma representação visual adequada, pequenas diferenças entre os dados podem ser evidenciadas, proporcionando sua percepção imediata.

O fato de existirem bibliotecas de visualização em variadas linguagens de programação permite que sejam desenvolvidos sistemas específicos para os dados que se deseja visualizar. Isso pode ser uma alternativa quando não se encontra uma ferramenta de visualização adequada, ou quando a intenção é inserir no sistema características ou funcionalidades específicas para o domínio.

Entre as diversas técnicas de visualização apresentadas neste capítulo, a *Tree-Map* foi selecionada para analisar os dados do INEP, mesmo esses dados não sendo hierárquicos.

A variação de cores e tamanhos da *Tree-Map* evidencia características de cada tupla do conjunto de dados, além da técnica usar todo o espaço disponível na tela, usando apenas retângulos ou quadrados, que se encaixam um nos outros, sem necessitar de arestas, formas geográficas arredondadas ou até mesmo formas geográficas diferentes umas das outras, o que consome mais espaço na visualização. Essa característica favorece a visualização de grandes conjuntos de dados, como é o caso das avaliações educacionais do INEP, como é mostrado no Capítulo 7.

Neste capítulo foram apresentados breves conceitos sobre Visualização de Informações, assim como técnicas e ferramentas utilizadas. Alguns trabalhos que utilizam visualização também foram apresentados, possibilitando constatar que as técnicas de visualização, além de serem amplamente utilizadas, estão em constante aprimoramento e são úteis para todas as áreas.

Capítulo 4

ONTOLOGIAS

Este capítulo apresenta conceitos sobre Ontologia, assim como métodos para seu desenvolvimento, tecnologias de apoio, desafios, pesquisas relacionadas e sua importância para a Web Semântica.

4.1 Considerações Iniciais

O Dicionário Oxford de Filosofia define ontologia como “o termo derivado da palavra grega “*onto*”, que significa “*ser*”, mas usado desde o século XVII para denominar o ramo da metafísica que diz respeito àquilo que existe” (BLACKBURN, 1997).

O estudo de ontologias na computação caracteriza-se como um ramo de pesquisa que surgiu no final dos anos 80, propondo alternativas para representar o conhecimento. Tornou-se também alvo de pesquisas frequentes na área de Ciência de Informação, pois permite compartilhar e utilizar, em diferentes contextos, o conhecimento obtido com ajuda de especialistas de um certo domínio, sendo muito útil, por exemplo, em tarefas de adaptação de conteúdo para diferentes dispositivos.

Atualmente, considerando o intenso uso da computação por intermédio da internet, percebe-se que informações substanciais para suporte a decisão são disponibilizadas na internet e que esse fato tem sido alvo de várias pesquisas, como é o caso das iniciativas no contexto da Web Semântica, que considera Ontologia como um dos alicerces para a tendência do uso de grande volume de dados por meio de sistemas Web.

Este capítulo apresenta na Seção 4.2 alguns conceitos sobre Ontologia, assim como sua utilidade e relevância para a Web Semântica. Na Seção 4.3 são apresentados métodos para o desenvolvimento de ontologias, com ênfase ao método Methontology, no qual esta pesquisa se apoia. Na Seção 4.4 algumas ferramentas e linguagens são mencionadas e na Seção 4.5, algumas pesquisas relacionadas são apresentadas, permitindo evidenciar como esta é uma área vasta e com muitas colaborações a serem feitas. Na Seção 4.6 apresentam-se as considerações finais sobre o capítulo.

4.2 Ontologias

A definição mais conhecida para o termo Ontologia, no contexto da computação, é de Gruber (1993, 1995), que diz que uma ontologia é a especificação formal e explícita de uma conceitualização compartilhada. Entretanto, para a correta interpretação desse conceito, Fensel (2001) esclarece que:

- **Conceitualização:** é uma visão simplificada e abstrata do domínio que se deseja representar. Portanto, são os objetos, os conceitos e outras entidades existentes em alguma área de interesse e os relacionamentos existentes entre eles;
- **Explícita:** indica que o tipo dos conceitos e as restrições disponibilizadas devem ser definidos de maneira explícita;
- **Formal:** mencionando o fato de que todas as ontologias devem ser codificadas em uma linguagem formal para evitar ambiguidades e serem passíveis de processamento automático;
- **Compartilhada:** referindo-se ao fato que uma ontologia capta o conhecimento consensual, não se restringindo a um indivíduo e sim, aceito por um grupo.

Os componentes principais de uma ontologia são:

- Classes: são representações concretas de um conceito e reúne indivíduos que satisfaçam seu conjunto de requisitos. O termo *conceito* também é utilizado para denominar classes;
- Indivíduos: também referenciados como *instâncias* ou *instâncias de classes*, representam os objetos no domínio para o qual a ontologia está sendo desenvolvida;
- Propriedades: são relações entre dois indivíduos ou, dependendo do tipo, duas classes. Também são conhecidas como *relações* em abordagens Orientadas a Objetos.
- Axiomas: são sentenças sempre verdadeiras que indicam restrições e inferências de conhecimento do domínio.

O uso de ontologias oferece vantagens como: melhoria da comunicação entre as pessoas envolvidas, afinal, facilita a obtenção de um consenso sobre o vocabulário e significado dos termos definidos; possibilita a formalização do conhecimento, evitando com isso ambiguidades e inconsistências; representa o conhecimento de um domínio, facilitando sua disseminação e reuso, pois descrevem o conhecimento de forma explícita, e também permite que, durante o desenvolvimento de diferentes aplicações para o domínio, haja o aperfeiçoamento do conhecimento, levando em consideração que essas aplicações podem ser desenvolvidas por equipes diferentes, em momentos distintos e com variadas finalidades.

Por representarem a semântica de um domínio, as ontologias podem ser usadas por diversas aplicações, incluindo aplicações Web. Por esse motivo, é considerada um importante recurso para o estabelecimento da Web Semântica.

O objetivo da Web Semântica é desenvolver e disponibilizar tecnologias e padrões projetados para auxiliar máquinas a processarem as informações disponibilizadas na Web, possibilitando, assim, a descoberta de informações implícitas e valiosas, a integração de dados e até mesmo a automatização de algumas tarefas.

4.3 Desenvolvimento de ontologias

Uschold e King (1995) preveem alguns estágios para o desenvolvimento de ontologias, que devem ser contemplados por todos os métodos. O mapa mental na Figura 4.1 ilustra os estágios definidos.

Como é possível analisar, as etapas envolvidas no processo de desenvolvimento de ontologias se assemelham às etapas de desenvolvimento de software. Para Devedžić (2002), ao colocar a Engenharia de Ontologia no contexto de outras disciplinas, em especial a Engenharia de Software, muitas semelhanças e analogias podem surgir. Essas semelhanças permitem fazer ligações a fim de compreender lacunas, além de ver e conhecer conceitos e práticas com diferentes perspectivas.

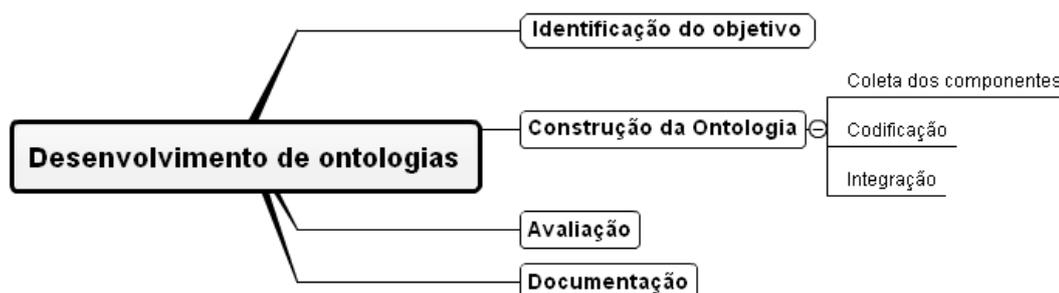


Figura 4.1. Estágios de desenvolvimento de ontologias de Uschold e King (1995).

A literatura apresenta diversos métodos utilizados para o desenvolvimento de ontologias. Segundo Gómez-Pérez, Fernández-López e Corcho (2003), desde 1990 vários pesquisadores publicaram pontos importantes para orientar a criação de ontologia, mas foi em 1995 que surgiram as primeiras propostas concretas, baseadas na experiência acumulada por pesquisadores no desenvolvimento da *Enterprise Ontology* (Uschold; King, 1995) e da *TOVE (TOronto Virtual Enterprise)* (Grüninger; Fox, 1995), propostas essas que foram refinadas nos anos seguintes.

Na mesma época, Schreiber, Wielinga e Jansweijer (1995) apresentaram o método usado para construção de uma ontologia no domínio das redes elétricas, como parte dos resultados do projeto *Esprit KACTUS*, do qual os autores participavam. Fernández, Gómez-Pérez e Juristo (1997) apresentaram o método *Methontology*, que ainda hoje é bastante propagado e utilizado por várias pesquisas.

Swartout e outros (1996) propuseram um novo método para a construção de ontologias baseado na ontologia SENSUS, que possui cerca de 70.000 conceitos sobre segurança eletrônica. Alguns anos mais tarde surgiu outro método, denominado On-To-Knowledge (Fensel; Harmelen, 1999), resultante de um projeto com o mesmo nome.

Gómez-Pérez, Fernández-López e Corcho (2003) apresentam um breve estudo sobre os métodos citados e concluem que nenhuma das abordagens é totalmente madura, principalmente se comparada com os métodos de Engenharia de Software. No entanto, o método Methontology, de acordo com os autores, se afirma como o mais completo, além de ser recomendado pela FIPA (*Foundation For Intelligent Physical Agents*) *Ontology Service Specification* para a tarefa de construção de ontologias.

Noy e McGuinness (2001), por meio do conhecido “*Ontology Development 101: A Guide to Creating Your First Ontology*” apresentam questões gerais de análise para desenvolvimento de ontologias e propõem um processo iterativo, composto por sete passos: 1) determinar o escopo da ontologia; 2) considerar o reúso; 3) listar termos; 4) definir classes; 5) definir propriedades; 6) definir restrições e 7) criar instâncias.

Para as autoras, o desenvolvimento de uma ontologia inclui basicamente:

- definição das classes da ontologia;
- organização das classes hierarquicamente;
- definição dos relacionamentos e descrição de seus valores permitidos; e
- preenchimento de valores dos relacionamentos para instâncias.

Muitos trabalhos ainda hoje são propostos, cada um com intuítos diferentes e destinados a uma ou algumas etapas específicas do processo de desenvolvimento de ontologias, sendo que estas podem variar de autor para autor.

Este trabalho foi influenciado pelo método Methontology, que como dito anteriormente, é considerado um dos mais completos e prevê diversas atividades para o desenvolvimento de ontologias, sendo a seção seguinte reservada para a sua apresentação.

4.3.1 O método Methontology

Desenvolvido no Laboratório de Inteligência Artificial da Universidade Técnica de Madrid (UPM), o método Methontology possibilita a criação de ontologias desde o início, apoia processos de reengenharia de ontologia e auxilia a criação com reutilização de outras ontologias (GÓMEZ-PÉREZ et al., 2003).

Na Figura 4.2 é ilustrado o método Methontology, na qual é possível observar todas as atividades previstas pelo método para desenvolvimento de ontologias, sendo que elas são divididas em Atividades de Gestão, de Desenvolvimento e de Apoio, cada qual correspondendo a um retângulo da figura.

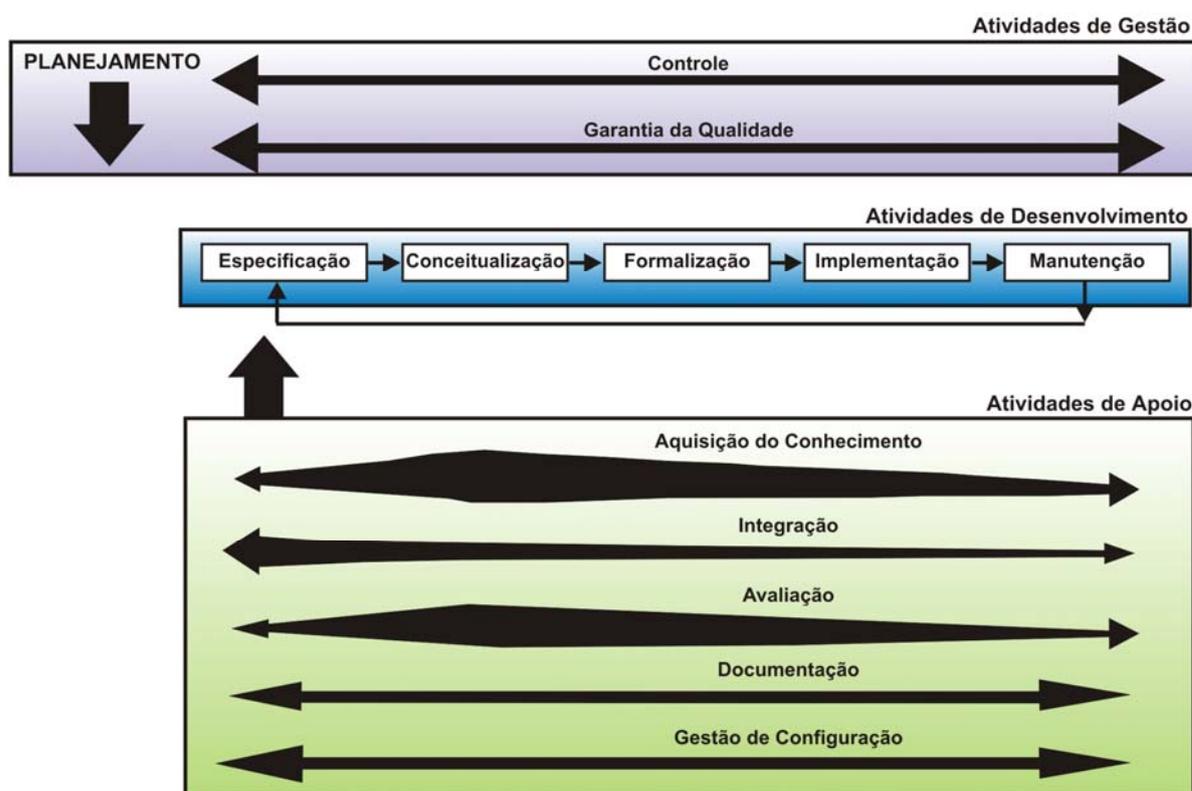


Figura 4.2. O método Methontology. Adaptado de Gómez-Pérez, Fernández-López e Corcho (2003).

As *Atividades de Gestão*, representadas no retângulo no topo da Figura 4.2, são: Planejamento, Controle e Garantia da Qualidade.

Baseando-se no relato de Fernández-López e outros (1999), o primeiro passo para construir uma ontologia, como em qualquer projeto, é elaborar um Planejamento, que deve ser controlado e deve conter atividades para Garantia da Qualidade.

Na literatura não foram encontrados critérios específicos sobre as Atividades de Gestão do método Methontology. Presume-se que na atividade de Planejamento, os objetivos do projeto sejam esclarecidos e responsabilidades sejam delegadas, além de definir entre os envolvidos atividades para efetuar o controle do desenvolvimento e para garantir a qualidade do produto final e dos artefatos gerados durante o processo.

As *Atividades de Desenvolvimento*, representadas no retângulo do centro da Figura 4.2, compõem o ciclo de vida do método, o qual é subdividido nas seguintes fases: Especificação, Conceitualização, Formalização, Implementação e Manutenção. Durante a realização dessas fases, a ontologia que está sendo construída pode sofrer alterações caso esteja utilizando outras ontologias e estas sofrerem alterações durante a construção da ontologia em questão.

O intuito da fase de *Especificação* da ontologia é criar um documento em linguagem natural, que abranja o objetivo principal da ontologia, sua finalidade, nível de granularidade e de alcance. Essa especificação deve ser tão completa quanto possível e concisa.

Na fase de *Conceitualização*, os conhecimentos não estruturados adquiridos nas fases anteriores devem ser organizados. Essa fase converte, informalmente, as informações do domínio em estudo em uma especificação semi-formal, utilizando um conjunto de representações intermediárias que possam ser compreendidas pelos especialistas do domínio e pelos desenvolvedores da ontologia. A tarefa inicial dessa fase é a construção de um glossário.

Para Gómez-Pérez, Fernández-López e Corcho (2003) essa é a principal fase no processo de desenvolvimento da ontologia na abordagem Methontology e é para essa fase de *Conceitualização* que este trabalho tem o objetivo de contribuir.

A fase de *Formalização* consiste em identificar e formalizar a criação dos componentes (classes, relações, axiomas e instâncias) da ontologia em desenvolvimento. Como a criação desses componentes pode ocorrer já na fase de *Conceitualização*, para Gómez-Pérez, Fernández-López e Corcho (2003) essa fase não é obrigatória.

De acordo com Gruber (1993) as classes da ontologia, abstraídas na fase de *Conceitualização*, são organizadas em uma taxonomia; as relações representam o tipo de interação entre os conceitos de um domínio; os axiomas são usados para modelar sentenças sempre verdadeiras e as instâncias são utilizadas para

representar elementos específicos, ou seja, os próprios dados do domínio para o qual a ontologia está sendo criada.

A fase de *Implementação* de ontologias requer a utilização de ambientes capazes de dar suporte às características das meta-ontologias selecionadas na atividade de Integração, descrita a seguir.

O resultado da fase de Implementação é a ontologia codificada em uma linguagem como a OWL (OWL, 2009) ou RDF (LASSILA; SWICK, 1998). Nessa etapa, são escolhidas a ferramenta e a linguagem para construção da ontologia. Por se tratar de uma tarefa dispendiosa, qualquer apoio na construção de ontologias pode representar ganhos significativos. Na Seção 4.3 são apresentadas ferramentas e linguagens comumente utilizadas.

Na fase de *Manutenção* a ontologia deve, quando necessário, receber melhorias ou reparos em possíveis defeitos. Ao observar a Figura 4.2, nota-se que a fase de Manutenção leva novamente à Fase de Especificação, indicando que sempre que houver a necessidade de uma manutenção na ontologia, seja para aumentar ou diminuir sua abrangência ou corrigir defeitos, todas as fases anteriores devem se repetir.

Esse ciclo de vida é amparado pelas *Atividades de Apoio*, que estão representadas no retângulo inferior da Figura 4.2, e são: Aquisição do Conhecimento, Integração, Avaliação, Documentação e Gestão de Configuração. De acordo com essa figura é possível observar que algumas tarefas são enfatizadas durante algumas fases, enquanto a execução de outras tarefas permanece constante. Essa característica pode ser observada por meio da espessura das setas que as representam.

Nota-se, por exemplo, que a atividade Aquisição de Conhecimento é enfatizada durante a fase de Conceitualização da ontologia, enquanto que atividades como Documentação e Gestão de Configuração permanecem constantes durante todo o processo.

A atividade de *Aquisição do Conhecimento* é uma atividade independente no processo de desenvolvimento da ontologia. Fernández-López e outros (1999) explicam que a aquisição de conhecimento normalmente diminui conforme evolui o desenvolvimento, pois é natural que os pesquisadores tornem-se mais familiarizados com o domínio. Os autores dividem a atividade de Aquisição de Conhecimento em três possíveis etapas:

- Reuniões preliminares com especialistas para obter um conhecimento geral sobre o domínio;
- Estudo da documentação obtida nas reuniões para abstrair ao máximo os ensinamentos transmitidos pelos especialistas;
- Análise progressiva do conhecimento geral obtido com o intuito de refinar e detalhar o conhecimento.

Os autores ainda sugerem o uso de técnicas de sistemas baseados em conhecimento - KBS (*Knowledge-Based System*), como por exemplo: entrevistas estruturadas e não estruturadas com especialistas, análise de texto formal e informal, tabelas e gráficos de análise do domínio, análise de unidades de medição e revisões detalhadas de um especialista nas informações produzidas.

A atividade de *Integração* é quando a equipe deve verificar se há possibilidade de reutilizar ontologias já criadas. No repositório de ontologias DAML (<http://www.daml.org/ontologies>), atualmente encontram-se mais de 280 ontologias.

A atividade de *Avaliação* diz respeito a julgar tecnicamente a ontologia, verificando se a implementação (quando estiver sendo feita) e a documentação estão de acordo com o que foi planejado na fase de Especificação. A Avaliação consiste em duas atividades: Verificação e Validação.

A Verificação consiste no processo técnico que garante a consistência da ontologia, do ambiente de software associado e da documentação ao longo do ciclo de vida da ontologia. A Validação garante que a ontologia corresponda ao que foi proposto representar (FERNÁNDEZ; GOMÉZ-PÉREZ; JURISTO, 1997).

Vale ressaltar que é importante que a atividade de Avaliação seja realizada em todas as fases do ciclo de vida, como pode ser observado na Figura 4.2.

De acordo com diversos autores, não há um consenso sobre a atividade de Documentação (FERNÁNDEZ; GOMÉZ-PÉREZ; JURISTO, 1997; ALMEIDA; BAX, 2003).

Fernandez e outros (1997) já afirmavam que se a intenção é fazer com que a ontologia desenvolvida seja reutilizada, é fundamental se preocupar em documentar todo o processo da melhor maneira possível. No método Methontology é recomendável obter um documento após cada fase do ciclo de vida. Entretanto, não

foram encontradas na literatura definições específicas e detalhadas para documentar cada fase.

A mesma lacuna encontrada na literatura a respeito da fase de Documentação também ocorre para a fase de Gestão de Configuração. Considera-se que o ideal é se basear nas atividades previstas na Engenharia de Software, com o intuito de gerir as modificações, identificar os produtos que podem ser modificados, estabelecer relacionamento entre eles e definir mecanismos para administrar as diferentes versões dos artefatos.

4.4 Ferramentas e linguagens para construção de ontologias

Existem diversas ferramentas e linguagens destinadas a apoiar a criação de ontologias. Como exemplo de ferramentas, baseando-se em Almeida e Bax (2003) e em Gómez-Pérez, Fernández-López e Corcho (2003), podem-se citar a OntoEdit, WebONTO, WebODE e a Protégé 2000.

A OntoEdit é um ambiente gráfico para edição de ontologias que permite inspeção, navegação, codificação e alteração. O modelo conceitual é armazenado usando um modelo de ontologia que pode ser mapeado em diferentes linguagens de representação. As ontologias são armazenadas em banco de dados relacionais e podem se implementadas em XML, FLogic, RDF (S) e DAML+OIL, linguagens que serão apresentadas adiante (SURE *et al.*, 2002).

WebONTO é uma ferramenta que possibilita a navegação, criação e edição de ontologias, representadas na linguagem OCML e disponibiliza uma biblioteca com outras ontologias. Permite o gerenciamento de ontologias por interface gráfica, inspeção de componentes, verificação da consistência da herança e trabalho cooperativo (DOMINGUE, 1998).

O WebODE é um ambiente para engenharia de ontologias, baseado em um servidor de aplicações, desenvolvido em formulários HTML e *applets* Java, que dá suporte à maioria das atividades de desenvolvimento de ontologias. A integração com outros sistemas é possível, importando e exportando ontologias codificadas em linguagem de marcação (ARPÍREZ *et al.*, 2001).

Por fim, a Protégé-2000 é um ambiente interativo para projeto de ontologias, de código aberto, que oferece uma interface gráfica para edição de ontologias e uma arquitetura para a criação de ferramentas baseadas em conhecimento. Por ter uma arquitetura modulada, permite a inserção de novos recursos, adequando-o ao projeto e ao usuário (NOY; FERGERSON; MUSEN, 2000).

A maioria das ferramentas para criação de ontologias permite exportar e importar arquivos codificados em diferentes linguagens de representação de conhecimento, criando um bom grau de portabilidade entre elas.

De acordo com Gómez-Pérez, Fernández-López e Corcho (2003), no início da década de 90 foi criada uma série de linguagens para ontologias, baseadas em lógica de primeira ordem, frames (estruturas de dados de uma base de conhecimento) e lógica descritiva. Foi desenvolvido também um protocolo, chamado OKBC (*Open Knowledge Base Connectivity*), que assim como as linguagens, segue o paradigma de Representação de Conhecimento e permite o acesso a ontologias implementadas em diferentes linguagens. Como exemplo dessas primeiras linguagens, podem-se citar a CycL, Ontolingua, LOOM, OCML e a FLogic.

A expansão da Web conduziu a criação de novas linguagens de ontologia que explorassem suas características.

Essas linguagens são classificadas como linguagens de ontologia baseada na Web ou linguagens de marcação de ontologias, pois são baseadas em linguagens de marcação como HTML (*HyperText Markup Language*) e XML (*Extensible Markup Language*), que proporcionam a representação e o intercâmbio dos dados. Como exemplo dessas linguagens podem-se citar a SHOE, XOL, RDF e RDF-*Schema*, OIL, DAML+OIL e a OWL, utilizada nesta pesquisa. A Figura 4.3 ilustra o relacionamento entre essas linguagens.

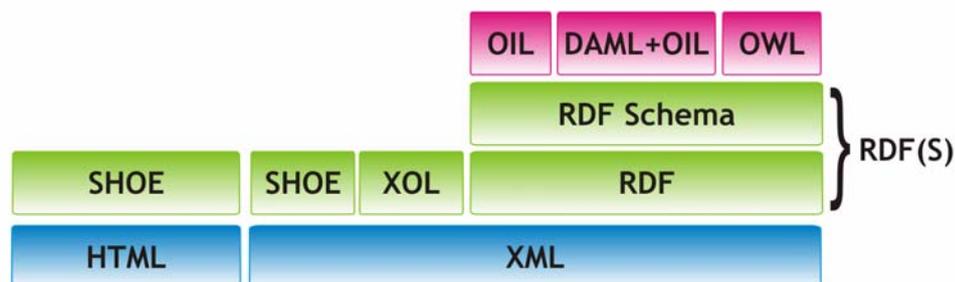


Figura 4.3. Relacionamento entre linguagens de ontologia baseada na Web. Adaptada de Gómez-Pérez, Fernández-López e Corcho (2003).

A linguagem SHOE (*Simple HTML Ontology Extension*) foi a primeira linguagem de ontologia baseada em Web a surgir. É uma extensão ao HTML, que permite adicionar marcações para inserir metadados em páginas Web. As marcações podem ser utilizadas para a construção de ontologias e para anotações em documentos Web. Como usa uma sintaxe SGML, pode ser usada dentro de documentos XML com apenas pequenas modificações (LUKE, HEFLIN, 2000).

A XOL (*XML Ontology Language*) é uma linguagem baseada em XML, que permite especificar conceitos, taxonomias e relações binárias, porém não possui mecanismos de inferência. A semântica da XOL é baseada em OKBC-Lite, que é uma forma simplificada do protocolo OKBC. Inicialmente, foi projetada para o intercâmbio de ontologias da área biomédica (KARP; CHAUDHRI; THOMERE, 2000).

A RDF (*Resource Description Framework*) foi desenvolvida pela W3C (*World Wide Web Consortium*) para descrever recursos Web, permitindo a especificação da semântica de dados baseada em um XML padronizado. Fornece também mecanismos para representar explicitamente serviços, processos e modelos de negócio, permitindo ao mesmo tempo reconhecimento de informações não explícitas (LASSILA; SWICK, 1998). A *RDF-Schema* também foi desenvolvida pela W3C, sendo uma extensão para a RDF, baseada em *frames*. Comumente, as duas linguagens são chamadas de RDF(S), sendo que ambas tornaram-se base para o desenvolvimento de outras linguagens, como a OIL, DAML+OIL e a OWL (GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ; CORCHO, 2003).

A OIL (*Ontology Interchange Language*) combina primitivas de representação de conhecimento baseada em *frames* com a semântica formal e com serviços de inferência da lógica descritiva. Da mesma forma que a OIL fornece uma extensão para o protocolo OKBC, também fornece uma extensão para as RDF(S). Por ser baseada na sintaxe das RDF(S), ontologias escritas em OIL são também documentos RDF válidos (HORROCKS *et al.*, 2000).

A DAML+OIL foi desenvolvida no projeto DAML da agência americana DARPA (*Defense Advanced Research Projects Agency*), baseada na especificação DAML (*Darpa Agent Markup :Language*) e na OIL. A DAML+OIL adiciona primitivas de representação de conhecimento baseada em lógica descritiva às RDF(S) (CONNOLLY *et al.*, 2001).

Em 2001 a W3C formou um grupo de trabalho nomeado WebOntology, com o objetivo de criar uma nova linguagem de marcação para ontologia baseada em Web Semântica, projetada para disponibilizar uma forma comum para o processamento de conteúdo semântico da informação na Web. O resultado desse grupo de trabalho foi a OWL – *Ontology Web Language* – que, conseqüentemente, é vista como uma importante tecnologia para a Web Semântica e é, atualmente, uma recomendação da W3C.

4.4.1 Características da OWL

De acordo com Horridge e outros (2004), dependendo da ontologia que se pretende criar, existem três variações da OWL que podem ser usadas:

- *OWL-Lite*: destinada a situações em que são necessárias apenas restrições e hierarquia de classes, sendo considerada sintaticamente mais simples;
- *OWL-DL (Description logic)*: mais expressiva, baseia-se em lógica descritiva, garantindo a computabilidade por meio de raciocínio automático, da verificação de inconsistências na ontologia e de decidibilidade;
- *OWL-Full*: mais expressiva que as anteriores, é destinada a situações em que a alta expressividade é mais importante do que garantir a decidibilidade ou completeza, o que faz com que não seja possível executar raciocínio automático nessa sub-linguagem.

Em OWL os componentes possuem diversas características que devem ser declaradas na ontologia, obedecendo as restrições de cada sub-linguagem.

Em relação às propriedades, em OWL existem dois tipos principais: Propriedades do Objeto, que relacionam um indivíduo a outro, e Propriedades de Tipo de Dados, que relacionam um indivíduo a um valor do XML *Schema* ou a um literal do RDF *Schema*. Exemplo: a propriedade “*temIrmã*” pode relacionar o indivíduo “*Ana*” ao indivíduo “*Maria*”, sendo essa uma Propriedade do Objeto. A propriedade “*temIdade*” pode relacionar o indivíduo “*Ana*” ao valor “*24*”, que é um tipo de “*xml: integer*”, sendo essa uma Propriedade de Tipo de Dados.

Uma Propriedade do Objeto pode ter uma Propriedade Inversa correspondente, fazendo com que se uma propriedade relaciona um indivíduo a outro, a propriedade inversa também relaciona inversamente esses indivíduos. Exemplo: declarando a propriedade “*temFilho*” como inversa de “*temPais*”, ao relacionar o indivíduo “*Maria*” ao indivíduo “*João*” com a propriedade “*temPais*”, automaticamente, “*João*” é relacionado ao indivíduo “*Maria*” por meio da propriedade “*temFilho*”.

Há também um terceiro tipo, nomeado de Propriedades de Anotação, que é usado para adicionar metadados às classes, aos indivíduos, às Propriedades do Objeto e às Propriedades de Tipo de Dados. Exemplo: a propriedade de anotação “*dc:creator*” pode relacionar a classe “*Receita*” ao dado literal “*Ana Maria*”.

Em OWL é possível enriquecer o significado das propriedades, por meio das Características das Propriedades, que podem ser funcionais, funcionais inversas, transitivas e simétricas.

As propriedades funcionais, também conhecidas como propriedades de valor único (*Single Value properties*) ou características (*Features*), garantem que para um indivíduo pode existir no máximo um indivíduo a ele relacionado. Exemplo: a propriedade “*temMaeBiologica*”, sendo declarada como funcional, garante que um indivíduo seja relacionado com essa propriedade a no máximo um indivíduo, pois uma pessoa possui apenas uma mãe biológica.

As propriedades funcionais inversas complementam as propriedades funcionais, fazendo com que se uma propriedade funcional conecta um indivíduo a outro, limitando-se ao número de um indivíduo, a propriedade funcional inversa também conecta, inversamente, esses indivíduos. Exemplo: uma propriedade funcional inversa para a propriedade “*temMaeBiologica*” pode ser a propriedade “*eMaeBiologicaDe*”, pois, se um indivíduo tem como mãe biológica outro indivíduo, indiscutivelmente, esse indivíduo é a mãe biológica do outro indivíduo.

As propriedades transitivas garantem que se um indivíduo está relacionado a outro indivíduo e este está relacionado a um terceiro indivíduo, significa que o primeiro indivíduo também está relacionado com o terceiro. Exemplo: se o indivíduo “*João*” está relacionado pela propriedade transitiva “*temAncestral*” ao indivíduo “*Pedro*” e “*Pedro*” está relacionado ao indivíduo “*José*” por meio da mesma propriedade, infere-se que o indivíduo “*João*” também está relacionado ao indivíduo “*José*”.

Se uma propriedade é transitiva ela não pode ser funcional, pois a funcional limita o relacionamento ao número de um indivíduo e a propriedade transitiva forma cadeias de indivíduos. Se uma propriedade é transitiva, a propriedade inversa a ela também se torna transitiva.

As propriedades simétricas garantem que se um indivíduo está relacionado a outro por meio desse tipo de propriedade, o relacionamento inverso entre esses indivíduos também é verdadeiro. Exemplo: se o indivíduo “Ana” se relaciona ao indivíduo “Maria” por meio da propriedade simétrica “*temIrmão*”, o indivíduo “Maria” também é relacionado automaticamente ao indivíduo “Ana” por meio dessa propriedade.

Além das peculiaridades descritas, as propriedades possuem Domínio (*Domain*) e Escopo (*Range*), o que significa que uma propriedade relaciona os indivíduos que pertencem à classe de Domínio dessa propriedade aos indivíduos que pertencem à classe de Escopo. Exemplo: a propriedade “*temAlunos*” pode ter como Domínio a classe “*Escola*” e como escopo a classe “*Alunos*”.

É possível definir mais de uma classe como Domínio ou como Escopo de uma propriedade. No entanto, interpretam-se essas classes definidas como Domínio ou como Escopo de uma mesma propriedade como uma união entre elas, o que em alguns casos, pode não ser verdade.

As classes em OWL podem ser declaradas como Classes Disjuntas, o que significa que um indivíduo não poderá ser ao mesmo tempo instância de classes declaradas disjuntas. Exemplo: se um indivíduo nomeado “Ana” for instância da classe “*Filha*”, e essa classe for declarada disjunta da classe “*Mãe*”, “Ana” não pode ser instância da classe “*Mãe*”, mas pode ser instância da classe “*Irmã*”, por exemplo. Ressalta-se que em OWL todas as classes são subclasses da classe “*Thing*”, que representa o conjunto de todos os indivíduos da ontologia.

As restrições são usadas para restringir os indivíduos de uma classe, sendo que as restrições podem ser divididas em três categorias: Restrições de Quantificador (*Quantifier Restrictions*), de Cardinalidade (*Cardinality Restrictions*) e Tem Valor (*hasValue*).

As Restrições de Quantificador são compostas por um quantificador, que pode ser existencial (\exists) ou universal (\forall), uma propriedade e um *filler*, que é a classe cujos indivíduos atendem à restrição.

Quando se usa o quantificador existencial, as restrições são chamadas de Restrições Existências (*someValuesFrom*) e descrevem o conjunto de indivíduos que tem pelo menos um tipo específico de relacionamento com indivíduos membros de uma classe.

Quando se usa o quantificador universal, as restrições são chamadas de Restrições Universais (*allValuesFrom*) e descrevem o conjunto de indivíduos os quais, para uma certa propriedade, são relacionados apenas com indivíduos da classe especificada no *filler*. Quando necessário, é possível combinar Restrições Existenciais e Universais.

As Restrições de Cardinalidade são compostas por um símbolo, uma propriedade e um número. O símbolo pode ser “*maior ou igual a*” (\geq), “*menor ou igual*” (\leq) e “*igual*” (=). Esse tipo de restrição é usado para especificar o número de indivíduos que podem participar de uma propriedade.

Quando se usa o símbolo “*maior ou igual a*” (\geq), as restrições são chamadas de Restrições de Cardinalidade Mínima, que especificam o número mínimo de relacionamentos que o indivíduo pode ter com a propriedade especificada.

Quando se usa o símbolo “*menor ou igual*” (\leq), as restrições são chamadas de Restrições de Cardinalidade Máxima, que especificam o número máximo de relacionamentos que o indivíduo pode ter com a propriedade especificada.

Quando se usa o símbolo “*igual*” (=), as restrições são chamadas de Restrições de Cardinalidade Exata, que especificam o número exato de relacionamentos que o indivíduo pode ter com a propriedade especificada.

Como se pode notar pelo que foi descrito nesta seção, há várias especificidades que podem ser declaradas em uma ontologia OWL. Ao selecionar uma ferramenta para desenvolver a ontologia, é aconselhável observar se a ferramenta contempla todas as especificidades e o quão fácil é criá-las para representar a ontologia que está sendo construída.

4.5 Trabalhos relacionados

Muitas pesquisas com o objetivo de propor novos métodos ou então complementar métodos propostos para o desenvolvimento de ontologias têm sido realizadas.

Falbo, Menezes e Rocha (1998) propõem um método sistemático para a construção de ontologias, aliado a uma linguagem gráfica de representação nomeada de GLEO (*Graphical Language for Expressing Ontologies*). O método apresentado pelos autores é composto das atividades de definição do propósito da ontologia e especificação de requisitos; captura da ontologia (equivalente à fase de conceitualização do método Methontology); formalização; integração com outras ontologias (quando necessário); avaliação; e documentação. Nota-se que as atividades definidas por esses autores podem ser relacionadas com atividades e fases do método Methontology, o que indica que a pesquisa apresentada nesta dissertação pode contribuir também com outros métodos.

Os autores sugerem que a documentação deve conter os requisitos, motivação, cenários, descrições textuais da conceitualização, os critérios adotados no projeto e um dicionário, composto por termos definidos na ontologia. Esse dicionário deve obedecer ao princípio da auto-referência, que indica que os termos usados para compor a definição de um termo devem, preferencialmente, serem termos descritos no dicionário. Além disso, deve ser obedecido o princípio do vocabulário mínimo, que indica que o vocabulário deve ser tão pequeno quanto possível e não deve apresentar ambiguidades. Os autores indicam o conceito de hipertexto para realizar essa documentação, pela possibilidade de navegação no documento. Essa característica é nativa no ambiente Web Moodle, que é usado no método proposto nesta pesquisa.

Holsapple e Joshi (2002) apresentam algumas abordagens para desenvolvimento de ontologia (Inspirativa, Indutiva, Dedutiva, Sintética e Colaborativa) e apresentam um método de desenvolvimento colaborativo.

Considerando a abordagem Inspirativa, os autores ressaltam que a ontologia resultante é baseada na criatividade, imaginação e ponto de vista individual do desenvolvedor, sendo muito útil, nesses casos, o uso de Sistemas de Apoio a Decisão (SAD) e Sistemas de Gestão de Informação (SGI). Salienta-se que produtos

resultantes dessa abordagem possuem uma utilização restrita, pelo fato de retratarem um ponto de vista individual, salvo os casos em que as impressões do desenvolvedor sobre o domínio em questão estão alinhadas com as impressões de outros possíveis usuários, ou quando o desenvolvedor tem a capacidade de impor o uso da ontologia.

Sobre a abordagem Indutiva, os autores explanam que a ontologia é desenvolvida por meio da observação, inspeção e análise de casos específicos no domínio de interesse. A caracterização da ontologia resultante de um dos casos pode ser aplicada em outros casos semelhantes. Por exemplo, uma ontologia de processos aplicados em uma empresa específica, desenvolvida por meio da abordagem Indutiva, pode ser aplicada em outras empresas.

Já a abordagem Dedutiva para construção de ontologias concentra-se em adotar e adaptar alguns princípios gerais de um domínio e aplicá-los para construir uma ontologia dirigida para um caso específico. Isso envolve a filtragem das informações gerais e sua customização para o domínio que corresponde a um subconjunto do domínio maior.

Sobre a abordagem Colaborativa os autores explicam que o desenvolvimento é um esforço conjunto que reflete experiências e pontos de vista dos colaboradores, o que ajuda a enriquecer o conteúdo da ontologia. Por outro lado, a coordenação do processo de concepção da ontologia pode se tornar difícil, caso ocorra o envolvimento de muitas pessoas. O processo pode estar firmemente ancorado com uma ontologia inicial ou ter como ponto de partida informações não estruturadas obtidas através de uma discussão inicial. A fim de executar uma abordagem colaborativa, um mecanismo de consenso deve ser empregado.

O método apresentado pelos autores é composto por quatro fases e é baseado no método Delphi (Linstone, Turoff, 1975)

Na Fase 1 são definidos critérios para a concepção da ontologia, especificação do escopo e normas para a avaliação da ontologia, sendo que essas estão baseadas nos conceitos de Gruber (1995) (clareza, coerência, extensibilidade, codificação mínima e comprometimento ontológico mínimo).

Na Fase 2 é produzida uma ontologia para servir como uma âncora para centrar a atenção dos colaboradores.

Na Fase 3 é utilizado o conceito do método Delphi para que a opinião de todos os envolvidos seja reunida e aconteça o desenvolvimento da ontologia. O

método Delphi é uma técnica formal para coleta e integração de opiniões de um grupo de pessoas sobre determinado assunto. Cada participante, isoladamente, relata suas opiniões em um questionário definido e após isso, um líder prepara um documento que reúne as respostas de todos do grupo. Esse documento é repassado a todos os envolvidos, os quais analisam suas respostas iniciais com base no documento formulado, modificam e incrementam suas respostas em um novo questionário e repassam novamente ao líder, que continua essa interação enquanto julgar necessário e promove, em cada fase, uma convergência das respostas de todos.

Na Fase 4 a ontologia é explorada, sendo utilizada em vários contextos para verificar sua abrangência.

Uma característica dessa pesquisa em relação à pesquisa apresentada nesta dissertação é que ambas utilizam um processo iterativo para reunir as informações de um grupo de pessoas. No entanto, o processo definido por Holsapple e Joshi (2002) promove uma discussão, na Fase 3, com base em uma ontologia pré-definida, o que, no processo apresentado nesta dissertação, acontece com base em um glossário. Caso exista uma ontologia inicial sobre o domínio, o refinamento da mesma pode acelerar o processo de compreensão do domínio tratado. Por outro lado, ao elaborar um glossário, vários especialistas do domínio podem contribuir, tornando-o uma fonte rica de informação. Além disso, a manipulação de glossários pode ser mais amigável, caso os especialistas do domínio não estejam acostumados a interpretar e manipular as representações utilizadas para descrever ontologias.

Kishore, Zhang e Ramesh (2004) sugerem um processo baseado no modelo espiral para desenvolvimento de ontologias - *Helix-Spindle Model*. O modelo é motivado pelo desenvolvimento de software Incremental, que além de ser um conceito popular em engenharia de software, permite aumentar o grau de formalidade para o desenvolvimento de ontologias.

O processo Helix-Spindle consiste em três fases principais: Concepção, em que há a representação informal do conhecimento; Elaboração, na qual é construída uma representação semi-formal; e Definição, na qual o conhecimento é formalizado. As três fases fazem parte das duas atividades definidas no processo: Construção e Teste da Ontologia. Na atividade de Construção são utilizados conhecimentos teóricos (dedutivos) tanto para a execução das fases, como para a abstração do domínio. Na atividade de Teste, o conhecimento utilizado é pragmático (indutivo).

A construção da ontologia utiliza conteúdos teóricos sobre o domínio a fim de extrair os componentes da ontologia. Ao fim de cada fase, o produto a ela relacionado é verificado e, caso ocorram incoerências, a fase é repetida. A ontologia resultante ao fim das três fases é, então, testada para assegurar que ela esteja coerente, consistente e que contenha o vocabulário que represente o domínio. Caso sejam detectados problemas relacionados à coerência ou à extensibilidade, o desenvolvimento da ontologia retorna para a fase inicial e o processo reinicia. O vocabulário produzido é usado na atividade de Teste para modelagem de um ou mais *frameworks* do domínio em questão.

A realização de testes ao fim de cada fase garante que problemas no desenvolvimento da ontologia sejam identificados precocemente, evitando correções posteriores ou até mesmo um produto final fraco ou que contenha erros.

A execução das três fases previstas por Kishore, Zhang e Ramesh (2004), passa por duas etapas: uma de construção de ontologia (*building*) e outra para teste (*testing*), na qual o usuário segue *checkpoints* para confirmar se o artefato resultante de cada fase está correto. Assim como nesse trabalho, é possível observar no Capítulo 6 que a proposta apresentada nesta dissertação também possui alguns pontos de checagem que contribuem para a melhoria da qualidade na construção de ontologias.

As pesquisas apresentadas acima não determinam especificamente o uso de alguma ferramenta para auxiliar a construção de ontologias. No entanto, como menciona Maedche e Staab (2001), a construção manual de ontologias é um processo complexo, de alto custo e propenso a erros.

Para suprir essa lacuna, diversos trabalhos vêm sendo propostos para a construção semi-automática ou automática de ontologias, com o intuito de otimizar o processo de construção, adaptação e manutenção das mesmas, empregando alguma fonte de conhecimento. Essa área de pesquisa é nomeada de *Ontology Learning*.

No entanto, como pode ser observado em Suryanto e Compton (2001), Gómez-Pérez e Manzano-Mancho (2003), que apresentam vários métodos de construção automática de ontologias, Baségio (2006) e Ribeiro (2008), a maioria dos trabalhos faz o uso de corpus anotados, com informações semânticas, sintáticas e/ou morfológicas, ou utilizam recursos como a rede semântica WordNet (2009) ou a base de conhecimento léxico SemCor (2009), por exemplo.

Essas características geram limitações em relação aos domínios e/ou idiomas para se criar uma ontologia, pois exigem a adequação do corpus em marcações específicas, como por exemplo, o XCES (<http://www.xces.org/>), ou restringem as características da equipe, pois os envolvidos devem entender sobre os recursos utilizados e sobre o formato adequado dos dados. Por outro lado, embora traga limitações, o uso desses recursos facilita a tarefa de construção de uma ontologia.

Assim, com o que foi apresentado sobre ontologias e seu processo de construção, percebe-se que há diversos trabalhos na literatura que contribuem com a modelagem, definição de hierarquias e construção de ontologias de maneira automática e métodos para o desenvolvimento de ontologias.

Nota-se também que existem algumas atividades de construção de ontologias que estão presentes em todos os métodos, como por exemplo, atividades para abstração do conhecimento (conceitualização), avaliação e documentação. Isso pode ser percebido no método Methontology, para o qual o trabalho realizado nesta pesquisa teve o objetivo de contribuir e facilitar a realização da fase de Conceitualização.

4.6 Considerações Finais

Ontologias são úteis em contextos que exigem compartilhamento, reúso, organização e processamento de informação semântica. Porém, o processo de construção de ontologias é considerado uma das principais dificuldades relacionadas ao tema. Por este motivo, novas pesquisas continuam sendo realizadas.

Ressalta-se que a escolha do método a ser seguido é uma decisão dos construtores da ontologia, uma vez que deve ser levada em consideração a adequação do método ao contexto de desenvolvimento e o conhecimento da equipe. Destaca-se também que, independentemente do método, a fase de conceitualização está presente em todos eles, mesmo que realizada de forma diferente em cada um.

Como mencionado, embora existam vários métodos, ferramentas e linguagens para a construção de ontologias, algumas atividades desses processos são mais ou menos exploradas. No caso deste trabalho, identificou-se claramente a necessidade e contribuição de uma ontologia para avaliações educacionais do INEP, em decorrência do uso inapropriado de termos nos questionários.

Considerando o contexto em questão, a principal dificuldade a ser solucionada neste caso está situada na atividade de Conceitualização, para que as próprias pessoas do domínio tomem ciência da necessidade dessa padronização e para que a ontologia, após definida, possa contribuir para a melhoria da qualidade dos instrumentos de avaliação educacional do INEP.

Assim, unindo esse problema com o problema do uso das questões ao longo do tempo, este trabalho explora o uso da visualização para auxiliar na solução dos mesmos.

Particularmente no caso da construção da ontologia, o uso de visualização facilita a geração automática de um arquivo OWL representando as informações construídas durante o processo de construção da ontologia, em especial na fase de Conceitualização, descrita neste capítulo.

Conceitos sobre ontologia, assim como métodos, ferramentas, linguagens, vantagens na sua utilização e alguns trabalhos correlatos foram apresentados neste capítulo, dando subsídios a esta pesquisa, que é apresentada com mais detalhes nos capítulos seguintes.

Capítulo 5

OS PROCESSOS P-LIDE E ONTOP E AS FERRAMENTAS DE APOIO

Este capítulo apresenta os processos P-LIDE e ONTOP, resultados desta pesquisa, e as ferramentas SEV-Tool e ONTOP-Tool, que apóiam, respectivamente, suas aplicações. O objetivo do P-LIDE é definir uma Linguagem para Integração de Dados Educacionais do INEP, a LIDE. O objetivo do processo ONTOP é apoiar a conceitualização de ontologias. Ambos os processos fazem uso de visualização para facilitar a execução das tarefas.

5.1 Considerações Iniciais

Considerando a necessidade de criação da LIDE, descrita no Capítulo 2, o objetivo inicial desta pesquisa foi a de criar mecanismos para analisar os dados referentes às avaliações educacionais do INEP e identificar nesses dados quais questões são usadas em quais anos e em quais questionários.

Esse desafio salientou o problema da falta de padronização dos termos usados para compor as questões, indicando a necessidade de uma ontologia para o domínio, provendo a padronização de termos e também a representação formal do domínio de avaliações educacionais do INEP.

Como proposta para a criação da LIDE, considerando todas as atividades necessárias para sua criação, esta pesquisa apresenta o P-LIDE – Processo para definição da Linguagem para Integração de Dados Educacionais, que tem na SEV-Tool (*Search and Edition in Visualization*) um apoio computacional para permitir que

os dados referentes às avaliações do INEP sejam padronizados sintaticamente, sendo que o uso de visualização facilita essa tarefa.

Como visto no Capítulo 3, sobre ontologias, uma das fases ainda pouco formalizadas é a de conceitualização do domínio para o qual uma ontologia está sendo criada. Os trabalhos relacionados a essa atividade apontam para a importância do envolvimento das pessoas conhecedoras do domínio e para um processo interativo e iterativo para que essa atividade explore, da melhor forma possível, o domínio a ser caracterizado na ontologia. Dessa forma, o processo ONTOP (*ONTOlogy Process*), proposto neste trabalho, procura explorar essa lacuna e fornecer subsídios para facilitar essa atividade.

Com o uso da *ONTOP-Tool*, ferramenta desenvolvida para apoiar a execução do processo e que assim como a *SEV-Tool*, faz uso de recursos de visualização de informação, o ONTOP dá suporte à fase de conceitualização de ontologias, partindo de um glossário até chegar à geração de um arquivo do tipo OWL, contendo a conceitualização inicial da ontologia pretendida.

Além da *ONTOP-Tool*, outras ferramentas já consolidadas apoiam o processo ONTOP, como a *Protégé-2000*, para continuidade no desenvolvimento da ontologia, e o ambiente *Web Moodle*, do qual é utilizado o recurso de compartilhamento e controle de glossários. Apesar da utilização da *Protégé 2000* nesta pesquisa, outras ferramentas de edição e criação de ontologias que permitem importar arquivos OWL também podem ser usadas.

Assim, o trabalho é apresentado neste capítulo da seguinte forma: a Seção 5.2 apresenta o processo P-LIDE, detalhando seus passos. Na Seção 5.3 as funcionalidades da *SEV-Tool* são detalhadas. Na Seção 5.4 dá-se uma visão do processo ONTOP, apresentando os passos que o compõem. Na Seção 5.5 são detalhadas as funcionalidades da *ONTOP-Tool*, as quais estão fortemente enlaçadas com a execução do próprio processo. Por fim, na Seção 5.6, apresentam-se as considerações finais.

5.2 O P-LIDE e a SEV-Tool

O P-LIDE - Processo para geração da LIDE (Linguagem para Integração de Dados Educacionais) – é um processo cujos passos levam à geração dessa linguagem, que é base para o desenvolvimento da plataforma do projeto Web-PIDE. O processo prevê a solução para um dos problemas identificados no contexto desse projeto, mencionados no Capítulo 1: identificar os anos e as avaliações do INEP em que cada questão foi utilizada.

Como são várias ocorrências da mesma avaliação e há diversos tipos de avaliação, fazer esse mapeamento de forma totalmente manual não era possível, pois significaria, para cada questão, procurar em todas as provas já realizadas, se a questão estava presente.

Outra possibilidade seria fazer uma busca automatizada, opção esta que foi descartada, pois qualquer diferença sintática geraria uma resposta não desejada. Além disso, a questão poderia estar escrita de forma diferente, mas ter exatamente o mesmo significado, o que necessita de uma decisão humana para padronizá-las.

Assim, identificou-se na visualização um recurso facilitador para realização dessa tarefa. A primeira tentativa foi encontrar ferramentas livres que pudessem apoiar essa tomada de decisão. Embora a TreeMap (TREETMAP, 2009), mencionada no Capítulo 3, fosse uma forte candidata, ela não fornecia dois recursos essenciais para o tratamento das questões: a busca e a edição.

O recurso de busca é necessário para que se possam encontrar questões que possuam palavras ou uma sequência de palavras em comum, sugerindo a possibilidade de serem a mesma questão. O recurso de edição é necessário para que se possa alterar a sintaxe de uma questão, para que ela se torne idêntica a outras questões que compartilhem o mesmo significado.

Dessa forma, resolveu-se implementar a ferramenta *SEV-Tool*, que usa a técnica *Tree-Map* como metáfora de visual e fornece os recursos de busca e edição necessários. Suas funcionalidades foram desenvolvidas com o objetivo de apoiar a execução do P-LIDE.

A técnica *Tree-Map* foi selecionada, pois, embora seja indicada para dados hierárquicos, mostrou-se eficaz para visualizar os questionários. Por utilizar 100% do

espaço disponível na tela, permite a visualização de um grande conjunto de dados sem perda de espaço para arestas ou desenhos gráficos.

Outro fator que colaborou para a escolha dessa técnica foi o fato de ser possível, por meio da biblioteca Prefuse, definir a cor e o tamanho dos retângulos que formam a visualização *Tree-Map*, baseando-se nos dados a serem visualizados.

O P-LIDE, por ser apoiado pela *SEV-Tool*, permite trabalhar com um grande conjunto de dados e, por meio da metáfora visual gerada pela ferramenta, destacar dados com base em palavras chaves, editá-los e tê-los reagrupados automaticamente, como é detalhado a seguir. O objetivo é a padronização sintática dos dados (metadados do INEP), agrupando os que são semanticamente equivalentes. Feito isso, é possível gerar, automaticamente, a LIDE e uma lista de termos usados no conjunto de dados.

A Figura 5.1 ilustra o P- LIDE que é composto por dois passos, descritos a seguir. Na Seção 5.3 é detalhada as funcionalidades da *SEV-Tool* que apóiam a execução do Passo 2.

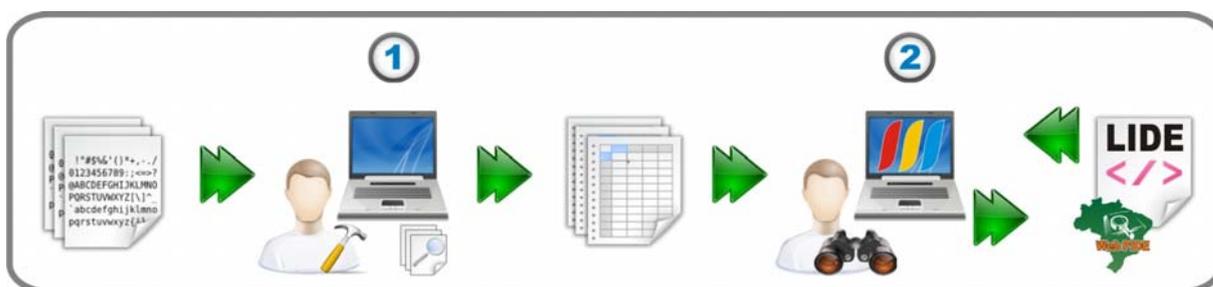


Figura 5.1. Processo P-LIDE.

- **Passo 1 – Preparar os dados para serem importados na *SEV-Tool*:** nesse passo é necessário executar as seguintes tarefas:
 - **Preparação dos dados para a padronização:** Como mencionado no Capítulo 2, o INEP disponibilizou diferentes tipos de arquivos com as informações sobre os questionários. Decidiu-se usar o arquivo SAS pela facilidade de manipulação e completude dos dados. Nesse passo os arquivos SAS são convertidos em texto (.txt).
 - **Padronização dos dados:** em alguns poucos casos os arquivos SAS não apresentam todas as informações necessárias sobre as questões. Nesses casos é necessário encontrar tal informação nos outros formatos de arquivos, como por exemplo o PDF disponibilizado pelo INEP. Durante a execução

dessa tarefa foi necessário um esforço para corrigir, por exemplo, erros de ortografia, a fim de tornar a execução do Passo 2 mais eficiente. Ao final dessa tarefa todos os arquivos SAS (convertidos em texto) devem estar compatíveis com o formato requerido na ferramenta *SEV-Tool*, descrito na seção seguinte.

- **Junção de questionários equivalentes:** depois da padronização no format dos questionários, todos os questionários equivalentes (de todos os anos) de todos os tipos de questionários (Português, Matemática, Biologia, etc.) devem ser reunidos no mesmo arquivo para que então possam ser utilizados na *SEV-Tool*.

- **Passo 2 – Gerar LIDE:** nesse passo é necessário executar as seguintes tarefas:
 - **Uso da *SEV-Tool* para analisar as questões de um questionário específico:** os arquivos formatados no Passo 1 devem ser importados na *SEV-Tool*. Baseado na visualização que a ferramenta constrói é necessário analisar as questões de todos os questionários que compõem o arquivo com o objetivo de corrigir a sintaxe de questões que são semanticamente iguais, mas estão sintaticamente diferentes. Quando o usuário é capaz de decidir pela alteração da sintaxe, é preciso decidir qual é a forma sintática mais adequada das questões e editá-las para que fiquem sintaticamente idênticas e sejam agrupadas na visualização gerada pela *SEV-Tool*. Essa tarefa, que deve ser repetida para todos os questionários (Português, Matemática, etc), é facilitada pelo recurso de visualização disponível na ferramenta, pois permite que o usuário identifique rapidamente questões isoladas e por meio do recurso de busca que a ferramenta oferece, permite encontrar questões de acordo com palavras-chave.
 - **Geração da LIDE:** Nesse momento, todos os arquivos que foram formatados no Passo 1 e padronizados sintaticamente na tarefa anterior devem estar no mesmo diretório. O usuário seleciona funcionalidade da *SEV-Tool* para geração da LIDE, indica o diretório que armazena os arquivos padronizados e é feita a geração do arquivo XML. O Capítulo 6 apresenta detalhes da geração desse arquivo.

Durante essa tarefa é possível gerar um outro arquivo XML, que contém os termos utilizados no conjunto de dados, sendo que termos comuns

da língua portuguesa (pronomes, artigos, etc) são desconsiderados. Com esse arquivo é possível construir um glossário no Moodle, construindo assim um relacionamento entre o processo P-LIDE e o processo ONTOP, que é descrito na Seção 5.4.

- **Simulação de buscas baseadas na LIDE:** após a geração da LIDE, a *SEV-Tool* permite que o usuário faça consultas no arquivo XML. Essas consultas simulam o uso da LIDE na plataforma Web-PIDE, a fim de responder questões básicas, como por exemplo, se há questões referentes a saneamento básico. Com essas consultas é possível saber se há questões referentes ao tema pesquisado, em que questionário a questão está e em quais anos ela foi utilizada. A seguir, essa e todas as funcionalidades da *SEV-Tool*, que apóiam a execução do Passo 2 do P-LIDE são detalhadas.

5.3 A SEV-Tool e suas funcionalidades

A *SEV-Tool* foi desenvolvida em Java, utilizando as bibliotecas gratuitas Prefuse (PREFUSE, 2009) para aplicar a técnica de visualização e a IDE (*Integrated Development Environment*) de desenvolvimento em Java NetBeans 6.5 (NETBEANS, 2009). É uma ferramenta *Desktop* e por isso, não necessita de recursos de rede para ser executada. Na Figura 5.2 é apresentada a tela de boas vindas da ferramenta.

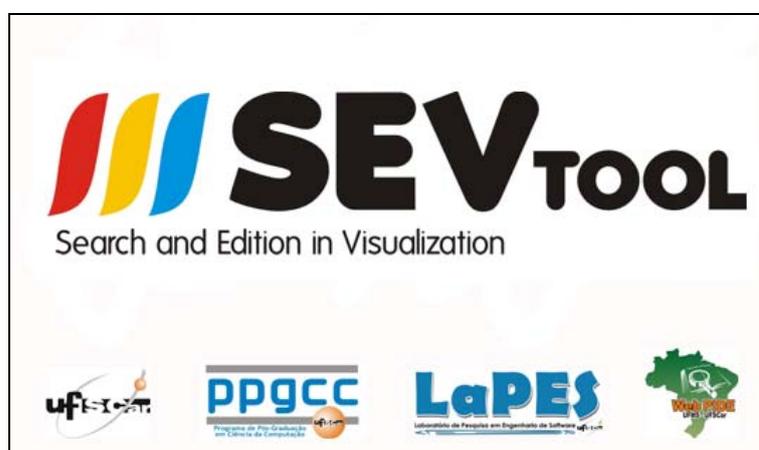


Figura 5.2. Tela de boas vindas da SEV-Tool.

A *SEV-Tool* não utiliza banco de dados, sendo que a persistência das informações definidas ou inseridas pelo usuário é feita em arquivos no formato texto (.txt).

A Figura 5.3 ilustra a tela inicial da ferramenta, na qual o usuário deve indicar qual é o diretório em que estão os arquivos que serão utilizados na execução do processo. A Figura 5.4 ilustra a tela na qual as funcionalidades da ferramenta, que apóiam a execução do Passo 2 do processo P-LIDE e suas respectivas tarefas, estão disponibilizados por meio dos botões.

Ressalta-se que o uso da ferramenta, assim como do processo *SEV-Tool*, é exemplificado detalhadamente no Capítulo 6.

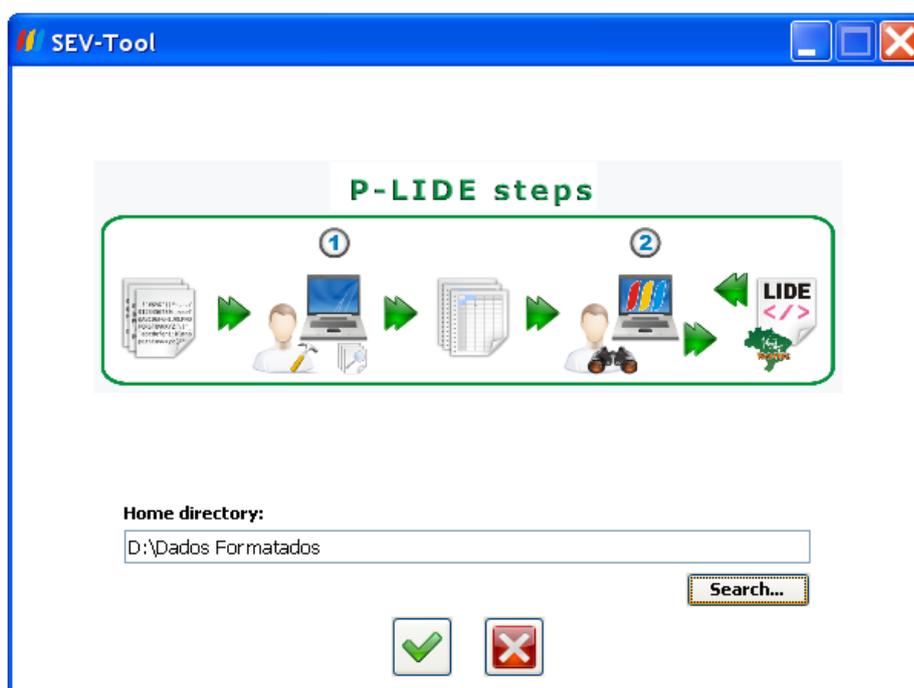


Figura 5.3. Tela inicial da *SEV-Tool*.

Em quase todas as telas, no canto superior direito, dois botões são disponibilizados para o usuário: um para voltar para a tela anterior (seta azul apontando para a esquerda) e outro para sair da aplicação (letra "X" em vermelho).

Por questão de organização, as funcionalidades providas nessa tela estão divididas em "*Questionnaire Analysis*" e "*Generate files with metadata*".

As funcionalidades agrupadas em "*Questionnaire Analysis*" possuem o mesmo objetivo: analisar e padronizar sintaticamente os metadados (questionários do INEP).

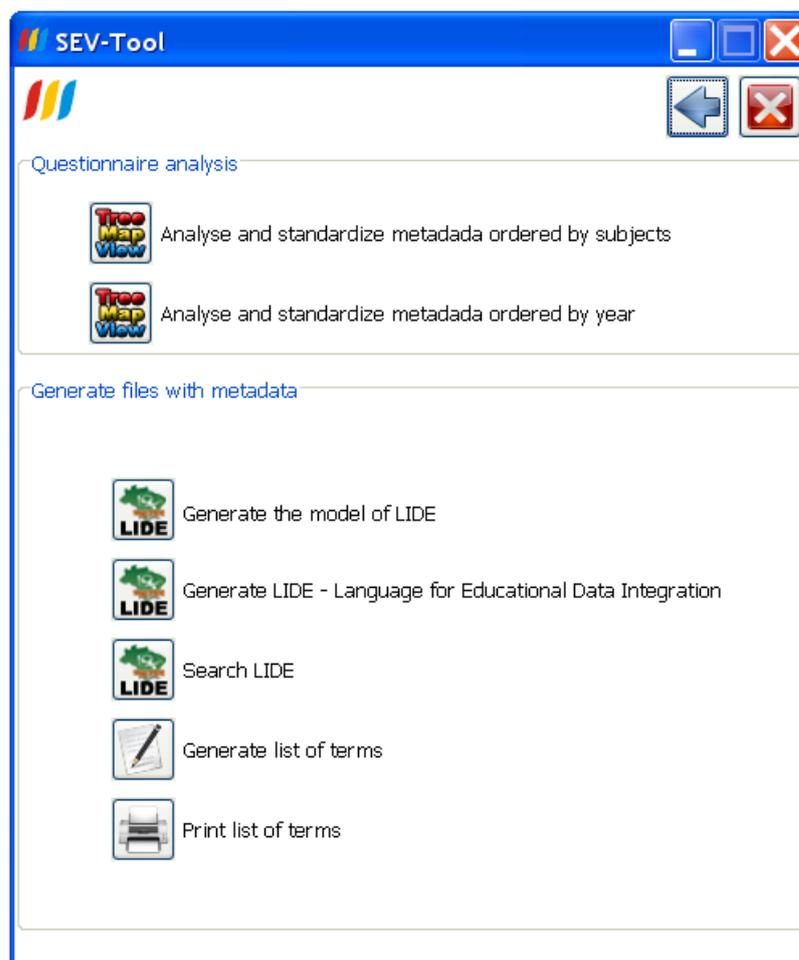


Figura 5.4. Tela com as funcionalidades da SEV-Tool.

Para visualizar o arquivo com um conjunto de questionários é necessário que o usuário clique no botão referente a esta funcionalidade na ferramenta (botão com o escrito *Tree-Map View*) e selecioná-lo.

Na versão atual, a *SEV-Tool* permite visualizar apenas os metadados das avaliações educacionais do INEP. Para que os questionários possam ser visualizados, é necessário que os metadados correspondentes estejam em formato texto (.txt) separados por tabulações e que respeitem os dados (atributos) previstos no arquivo, como ilustra a Figura 5.5.

Essa figura mostra um arquivo de uma planilha eletrônica o qual corresponde a um dos arquivos em formato texto que contêm os metadados de um dos questionários do INEP. A apresentação dos dados por meio da planilha permite a separação correta das tabulações do arquivo e facilita sua visualização neste texto. Os dados contidos no arquivo são, respectivamente, o nome do questionário ao qual a questão pertence, o ano que o questionário foi aplicado, a posição que o atributo

se inicia nos microdados do INEP, o nome do atributo no banco de dados do INEP, o tipo do atributo, um caractere “n” que permite à ferramenta identificar se a tupla foi alterada e, por fim, a descrição do atributo, que é a questão da avaliação.

PORTUGUES_04SERIE	1995	236	A041_003	numerico	n	Voce se considera:
PORTUGUES_04SERIE	1995	244	A041_004	numerico	n	Voce mora:
PORTUGUES_04SERIE	1995	252	A041_005	numerico	n	Em sua casa compra-se jornal?
PORTUGUES_04SERIE	1995	260	A041_006	numerico	n	Quantas horas por dia voce assiste a programas de televisao?
PORTUGUES_04SERIE	1995	268	A041_007	numerico	n	Voce costuma conversar em casa sobre o que acontece na escola?
PORTUGUES_04SERIE	1995	276	A041_008	numerico	n	Quantos dias voce faltou as aulas este ano?
PORTUGUES_04SERIE	1995	284	A041_009	numerico	n	Ha quantos anos voce esta no 1º Grau?
PORTUGUES_04SERIE	1995	292	A041_010	numerico	n	Voce teve que abandonar a escola durante algum tempo?
PORTUGUES_04SERIE	1995	300	A042_001	numerico	n	Voce achou o teste de Portugues:
PORTUGUES_04SERIE	1995	308	A042_002	numerico	n	Quanto tempo voce costuma gastar para fazer as lições de casa de Portugues?
PORTUGUES_04SERIE	1995	316	A042_003	numerico	n	Voce gosta de ler?
PORTUGUES_04SERIE	1995	324	A042_004	numerico	n	Voce usa jornais e revistas para fazer as lições de casa?
PORTUGUES_04SERIE	1995	332	A042_005	numerico	n	Seu professor passa leitura de livros como lição de casa?
PORTUGUES_08SERIE	1995	1	ANO	numerico	n	Ano
PORTUGUES_08SERIE	1995	9	SERIE	numerico	n	Serie
PORTUGUES_08SERIE	1995	11	DISC	alfanumerico	n	Disciplina
PORTUGUES_08SERIE	1995	12	TURMA	numerico	n	Numero da turma
PORTUGUES_08SERIE	1995	20	ALUNO	numerico	n	Codigo do aluno
PORTUGUES_08SERIE	1995	28	IDADE	numerico	n	Idade
PORTUGUES_08SERIE	1995	36	ESTRATO	numerico	n	ESTRATO
PORTUGUES_08SERIE	1995	44	UPA	numerico	n	Unidade Primaria de Amostragem
PORTUGUES_08SERIE	1995	52	DEP_ADM	numerico	n	Dependencia Administrativa(Estadual/Municipal/Particular)
PORTUGUES_08SERIE	1995	60	LOCAL	numerico	n	Localizacao(Urbano/Rural)
PORTUGUES_08SERIE	1995	68	REDE	numerico	n	Rede de Ensino(Publica/Particular)
PORTUGUES_08SERIE	1995	76	UF	numerico	n	Unidade da Federaçao
PORTUGUES_08SERIE	1995	84	REGIAO	numerico	n	Regiao
PORTUGUES_08SERIE	1995	92	PESO_AC	numerico	n	Peso calibrado(usado para expansao)

Figura 5.5. Formato de entrada para a visualização dos questionários na SEV-Tool.

O nome do questionário, presente na primeira coluna do arquivo, pode ser o nome de uma disciplina que possui um questionário na avaliação (português, matemática, biologia, etc), e quando é esse o caso, geralmente o nome do arquivo também é composto pela série escolar à qual o questionário foi destinado (4ª série, 8ª série ou 3º ano, referindo-se ao último ano do ensino médio); o nome da ocupação da pessoa que o respondeu (diretor, docente, etc) ou pode ser sobre o grupo de pessoas que o questionário avalia (turma).

A visualização dos questionários do INEP que seguem o padrão definido na ONTOP-Tool pode ser feita com os metadados organizados de acordo com o nome do questionário, ou com os metadados organizados por ano. Se o usuário deseja a primeira opção, todos os questionário (PORTUGUES_04SERIE, PORTUGUES_08SERIE, PORTUGUES_03ANO) de todos os anos que ele deseja visualizar devem estar no mesmo arquivo. Caso a segunda opção seja mais adequada, todos questionários de um determinado ano devem estar no mesmo arquivo (PORTUGUES_04SERIE, MATEMATICA_04SERIE, CIENCIAS_04SERIE, PORTUGUES_08SERIE, MATEMATICA_08SERIE, BIOLOGIA_08SERIE, etc).

Com base nos dados disponibilizados pelo INEP para o projeto Web-PIDE, todos os questionários se adaptam a esse formato. No entanto, sempre que o usuário desejar visualizar um conjunto de dados (metadados ou não), mesmo não sendo questionários do INEP, é necessário formatá-lo de acordo com o formato descrito acima.

É importante ressaltar que na visualização, os atributos são agrupados quando suas descrições (sétima coluna da Figura 5.5) forem sintaticamente idênticas e a cor dos atributos na metáfora visual é definida com base nos dados da terceira coluna, como pode ser observado a seguir.

Na Figura 5.6 é apresentada a tela da ferramenta para analisar e padronizar sintaticamente os metadados. Para efetuar a padronização, os questionários devem estar organizados de acordo com o nome (Português, Matemática, Docentes, etc, por exemplo).

Como a funcionalidade para visualização de questionário foi implementada para os questionários do INEP, antes que a *SEV-Tool* gere a visualização, duas tarefas são executadas automaticamente: a transformação das questões (descrição do atributo) em caracteres em minúsculo (caixa baixa) e a correção automática de algumas palavras que foram diversas vezes identificadas com erros de grafia ou falta de padronização. Essa correção foi necessária porque essas características impossibilitavam o agrupamento dos atributos mesmo quando as questões (descrição do atributo) eram idênticas. Um exemplo pode ser observado no trecho de texto "*rede de ensino (pública/particular)*", que em algumas questões aparece com o espaçamento diferente, como "*rede de ensino(publica/particular)*" ou "*rede de ensino (publica / particular)*".

Cada retângulo da metáfora visual representa uma questão, sendo que a cor é definida de acordo com o ano no qual cada questão foi utilizada. Questões idênticas são agrupadas automaticamente pela ferramenta, como é possível observar no canto superior esquerdo da Figura 5.6. Questões com sintaxe única ficam isoladas, como é possível observar no canto inferior direito da mesma figura.

Nota-se que na parte inferior da tela da Figura 5.6 é disponibilizado o recurso de busca (*search*) e informações sobre o conjunto de questionários que está sendo visualizado, como o total de questões, número de questões que foram editadas, número de questões presentes em todos os questionários e questões presentes em apenas um questionário.

Ao clicar em um retângulo da metáfora visual, as informações sobre a questão que ele representa podem ser visualizadas. Essas informações são: nome do atributo no banco de dados do INEP, tipo de dados do atributo, ano em que a questão foi aplicada, a qual questionário ela pertence e a descrição do atributo que, no caso dos metadados do INEP, é a questão direcionada ao aluno, professor ou diretor em cada avaliação.

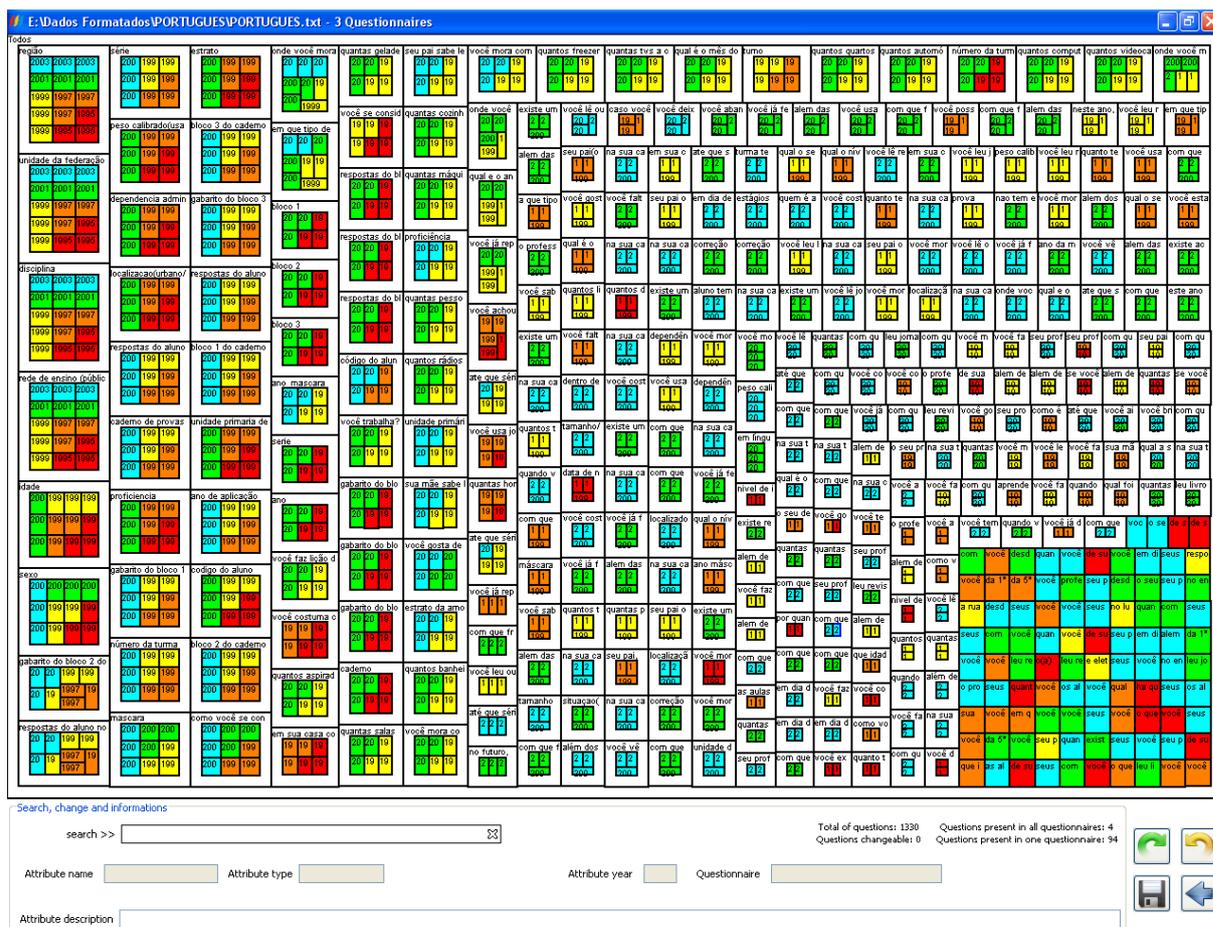


Figura 5.6. Tela da SEV-Tool para visualização dos questionários organizados por nome.

Para padronizar sintaticamente as questões, o usuário deve editar o campo de descrição do atributo (que é o único campo que pode ser editado) e salvar a alteração clicando no botão adequado (seta verde apontando para a direita no canto inferior direito). Imediatamente a ferramenta refaz a visualização, de forma que se a questão editada for igual a uma ou mais questões, estas serão visualizadas de forma agrupada.

Os retângulos que representam questões que foram editadas ficam com a coloração mais clara, o que permite ao usuário saber se a análise chegou ao fim. A

qualquer momento o usuário pode salvar o arquivo que está sendo visualizado e continuar a análise e padronização posteriormente.

A funcionalidade que permite visualizar os metadados organizados por ano foi implementada apenas para que o usuário possa analisar um conjunto de questionários e encontrar, de maneira mais rápida, as modificações entre os questionários de um mesmo ano. Na Figura 5.7 é apresentada a tela da ferramenta que permite essa visualização.

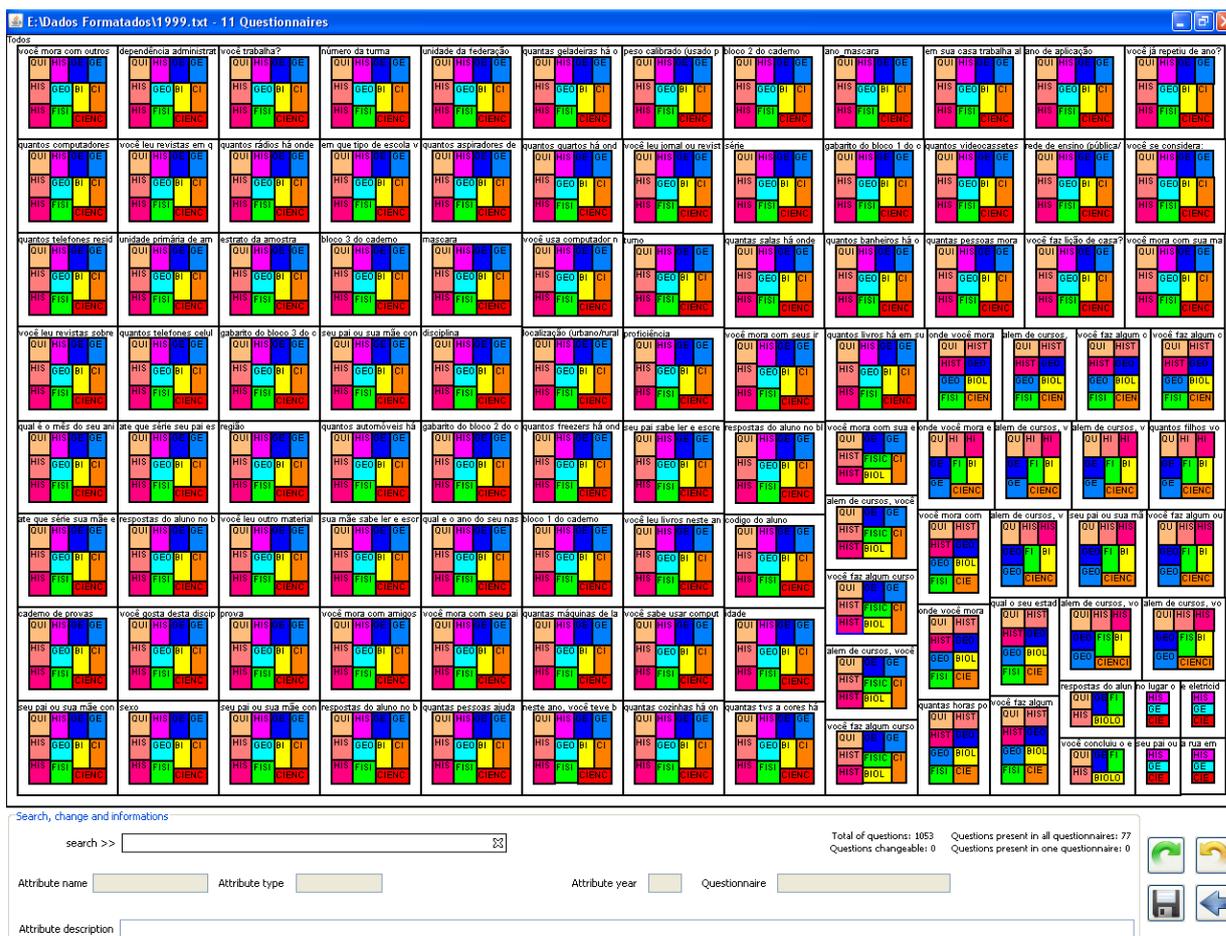


Figura 5.7. Tela da SEV-Tool para visualização dos questionários organizados por ano.

As funcionalidades agrupadas em “Generate files with metadata” objetivam a geração dos arquivos e a consulta à LIDE, como previsto no P-LIDE (Figura 5.1).

Com relação à definição da LIDE, o usuário pode gerar um arquivo com o modelo da LIDE, apenas para observar a estrutura do arquivo XML, ou então gerar a LIDE propriamente. Para a primeira opção é necessário clicar no botão com a descrição “Generate the model of LIDE” e para a segunda, “Generate LIDE – Language for Educational Data Integration”.

Nessa segunda opção, o usuário deve indicar o local em que deve ser salvo o arquivo XML e de qual avaliação é a LIDE que será gerada (SAEB, Censo Escolar, ENEM, Enade, etc), pois este será o nome do arquivo XML. A tela para inserção dessas informações pode ser vista na Figura 5.8.

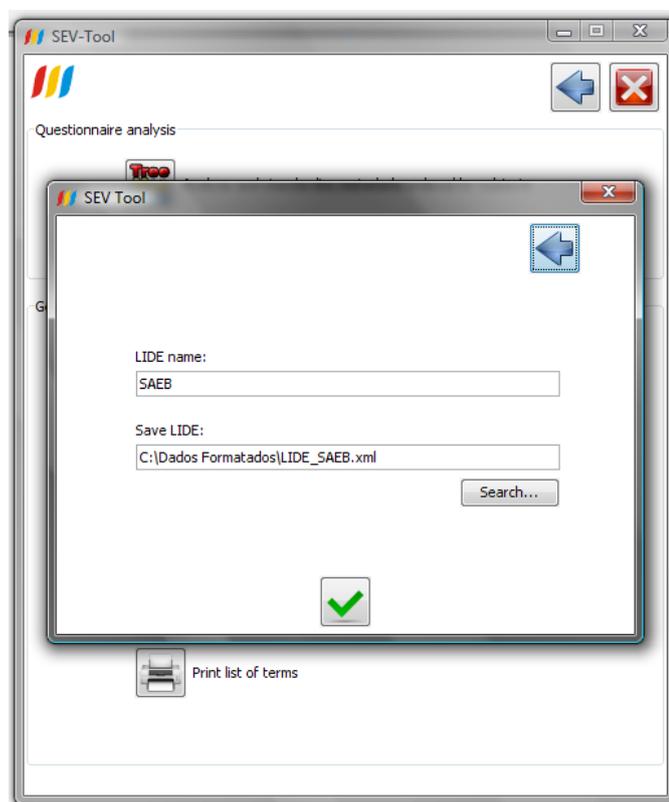


Figura 5.8. Tela da SEV-Tool para geração da LIDE.

Para gerar a lista de termos, a SEV-Tool usa uma lista de palavras que devem ser desconsideradas, lista esta comumente chamada de *stopwords*. Essa lista deve ficar na unidade raiz na máquina em que a SEV-Tool é executada (unidade C:, por exemplo). Essa lista pode ser incrementada e é formada por diversas palavras da língua portuguesa, como artigos e pronomes, que certamente não correspondem a entidades que caracterizam o domínio para o qual a ontologia está sendo criada.

Após a geração da LIDE, o usuário pode realizar consultas ao arquivo gerado, simulando o uso da linguagem pela plataforma Web-PIDE, como pode ser observado na Figura 5.9.

Sempre que novos questionários ou avaliações forem padronizados sintaticamente por meio do P-LIDE (e da SEV-Tool), uma nova LIDE deve ser gerada, tornando a linguagem cada vez mais completa.

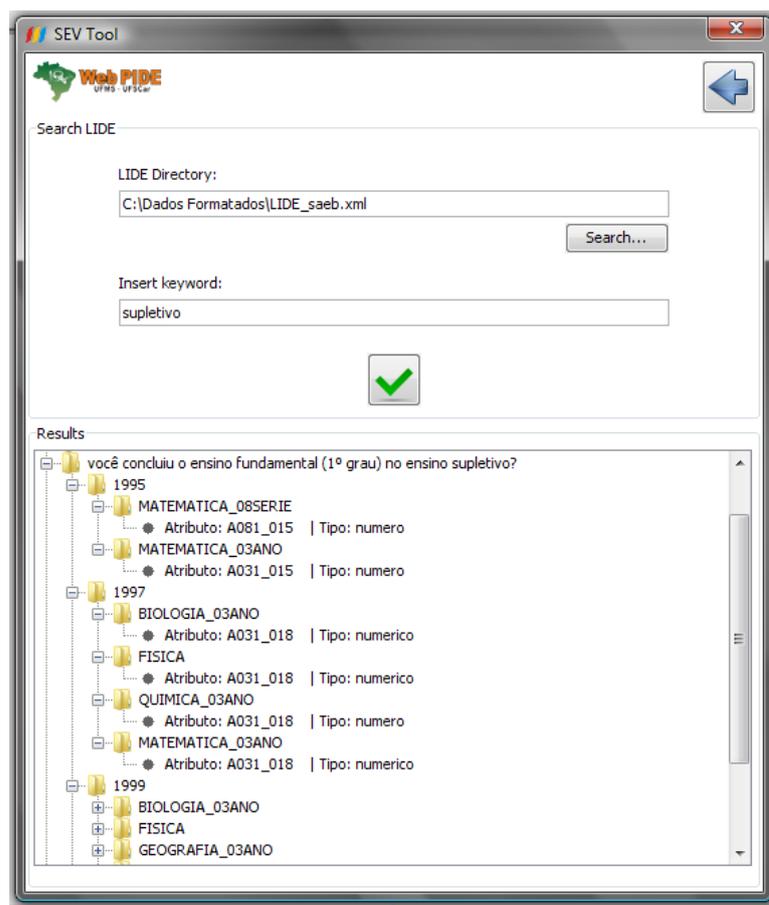


Figura 5.9. Tela da SEV-Tool para consultas na LIDE.

Para finalizar, a SEV-Tool disponibiliza a geração de uma lista com os termos utilizados no conjunto dos metadados (usados para a geração da LIDE). O usuário indica o diretório onde estão os arquivos base para a geração dessa lista e a SEV-Tool, automaticamente, faz a geração do arquivo XML em conformidade com o padrão do Moodle, para que este possa ser importado no ambiente como um glossário, criando um elo entre os processos P-LIDE e ONTOP, sendo este último descrito na próxima seção.

5.4 O processo ONTOP e a ONTOP-Tool

ONTOP é um processo que dá suporte à conceitualização de ontologias, partindo do refinamento de um glossário. Seus passos são executados com o apoio da ONTOP-Tool.

A criação do processo foi motivada pela necessidade de padronizar os termos usados para compor as questões das avaliações educacionais do INEP, detectada durante a geração da LIDE (processo P-LIDE). Após estudos na literatura, constatou-se a utilidade de uma ontologia para esse domínio e decidiu-se, então, criar o processo ONTOP. Por se tratar de passos genéricos, o processo não se restringe somente ao domínio do INEP.

O ONTOP reúne as atitudes tomadas e as soluções desenvolvidas no decorrer do desenvolvimento desta pesquisa em relação a conceitualização de ontologias. A cada novo desafio, novas decisões e funcionalidades foram adicionadas a ferramenta e, simultaneamente, o processo foi formalizado. Dessa forma, é notória a maneira como o processo e a ferramenta se fundem.

A motivação inicial para implementação da ferramenta *ONTOP-Tool* foi o desejo de criar um apoio computacional para a conceitualização inicial da ontologia sobre avaliações educacionais do INEP, objeto de estudo desta pesquisa.

O objetivo do processo é refinar um glossário que representa um domínio de aplicação e identificar, entre os termos, os possíveis componentes da ontologia que está sendo elaborada para esse domínio. Após a identificação dos componentes, é possível organizá-los de maneira hierárquica e criar alguns relacionamentos, para que após a execução desses passos, seja feita a geração de um arquivo OWL, que poderá ser importado em uma ferramenta específica de criação e edição de ontologias.

A Figura 5.10 representa visualmente esse processo, cujo os passos são explicados na sequencia.

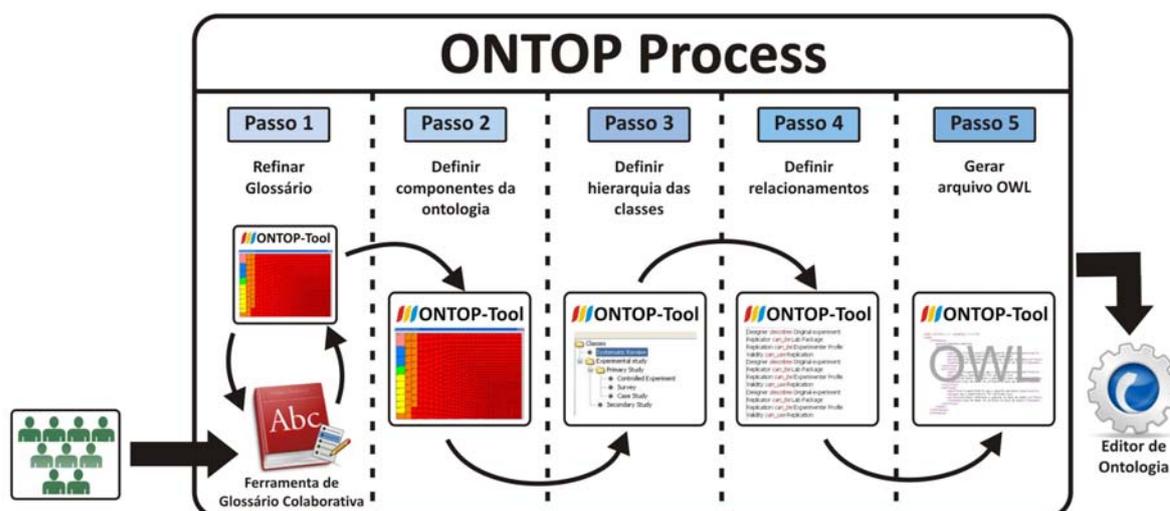


Figura 5.10. O processo ONTOP.

A entrada para o passo inicial é um glossário do Moodle, que pode ser resultante da execução do Processo P-LIDE (Seção 5.2) ou pode ser criado diretamente no ambiente Moodle.

A escolha do Moodle (*Modular Object-Oriented Dynamic Learning Environment*) ocorreu por ser uma ferramenta Web livre e disponibilizar a criação de glossários, com os recursos necessários para gerenciá-los. Com isso, especialistas do domínio podem colaborar com a construção do glossário, fazendo com que esse retrate fidedignamente, o domínio abordado.

O processo é composto por 5 passos:

- **Passo 1 – Refinar Glossário:** O objetivo desse passo é refinar os termos do glossário, para que esse contenha apenas os termos adequados ao domínio, e validá-los, verificando a coesão entre as descrições e se não há redundância de termos. Para isso, foi estabelecido um processo iterativo para que as ações sejam feitas, coletivamente, com a participação de especialistas no domínio. A iteratividade é realizada por meio da exportação/importação do glossário no Moodle e na *ONTOP-Tool*, até que os envolvidos admitam que o glossário representa o domínio em questão de maneira adequada;
- **Passo 2 – Definir componentes da Ontologia:** Nesse passo o objetivo é identificar, com o apoio da visualização e dos recursos disponibilizados por ela (cores e tamanhos de acordo com a frequência de cada termo), os possíveis componentes da ontologia dentre os termos do glossário. A *ONTOP-Tool* permite classificá-los como classe, indivíduo (instância de uma classe), propriedade (relação) ou sinônimo de outro termo;
- **Passo 3 – Definir hierarquia das Classes:** O objetivo desse passo é permitir que o usuário organize hierarquicamente as classes definidas no passo anterior. Com o apoio da *ONTOP-Tool* o usuário cria a hierarquia de classes e pode, simultaneamente, verificar se são adequadas à ontologia, além de abstrair possíveis relacionamentos entre elas;

- **Passo 4 – Definir relacionamentos:** O objetivo desse passo é definir relacionamentos entre as classes definidas no Passo 2. Com o apoio da *ONTOP-Tool*, relacionamentos do tipo *Domain-Range*, descritos no Capítulo 4, podem ser criados, utilizando propriedades pré-definidas na ferramenta, propriedades obtidas das informações geradas no Passo 2 e novas propriedades, inseridas pelo usuário quando necessário;
- **Passo 5 – Gerar arquivo OWL:** Para finalizar o processo, neste passo ocorre a geração do arquivo OWL, que contém todas as informações definidas pelo usuário até o Passo 4. Com isso, essa pré-conceitualização da ontologia pode ser importada em uma ferramenta para criação e edição de ontologias, como a Protégé-2000 e outras, citadas no Capítulo 4.

5.5 ONTOP-Tool e suas funcionalidades

Assim como para a tarefa de padronização sintática dos questionários do INEP, objetivo do processo P-LIDE descrito anteriormente, identificou-se na visualização um recurso facilitador para realização da conceitualização, pois facilitaria a análise simultânea de todos os termos do glossário. Como os recursos de busca e edição, necessário na *SEV-Tool* também seriam úteis nesse processo, optou-se em desenvolver uma nova ferramenta a usar ferramentas disponíveis para uso.

Como a *SEV-Tool*, a *ONTOP-Tool* usa a técnica *Tree-Map* como metáfora de visual. A ferramenta também foi desenvolvida em Java, utilizando as bibliotecas gratuitas Prefuse (PREFUSE, 2009) para aplicar a técnica de visualização, a biblioteca OWLAPI (OWLAPI, 2009), para gerar o arquivo OWL e a ferramenta IDE (*Integrated Development Environment*) de desenvolvimento em Java NetBeans 6.5 (NETBEANS, 2009). É uma ferramenta *Desktop* e por isso, não necessita de recursos de rede para ser executada.

A *ONTOP-Tool* não utiliza banco de dados, sendo que a persistência das informações definidas ou inseridas pelo usuário é feita em arquivos no formato texto

(.txt). Em alguns casos os arquivos são salvos com extensões diferentes, como por exemplo, “ontop”, para indicar em qual funcionalidade da ferramenta ele foi gerado.

Na Figura 5.11 é apresentada a tela de boas vindas da ferramenta.

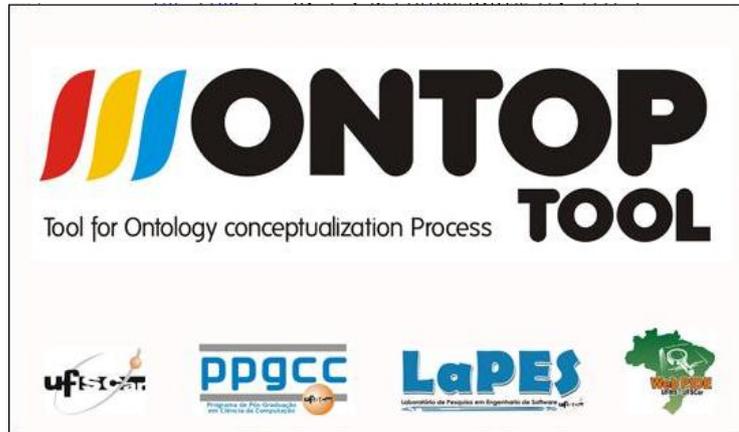


Figura 5.11. Tela de boas vindas da ONTOP-Tool.

A Figura 5.12 ilustra a tela inicial da ferramenta, na qual o usuário deve indicar qual é o diretório em que estão os arquivos que serão utilizados na execução do processo. Nessa tela, o usuário escolhe qual etapa do processo ele deseja iniciar. Como se pode observar, a tela inicial da ferramenta traz a visualização gráfica do processo ONTOP, o que reforça que a ferramenta foi moldada para o processo e vice-versa.

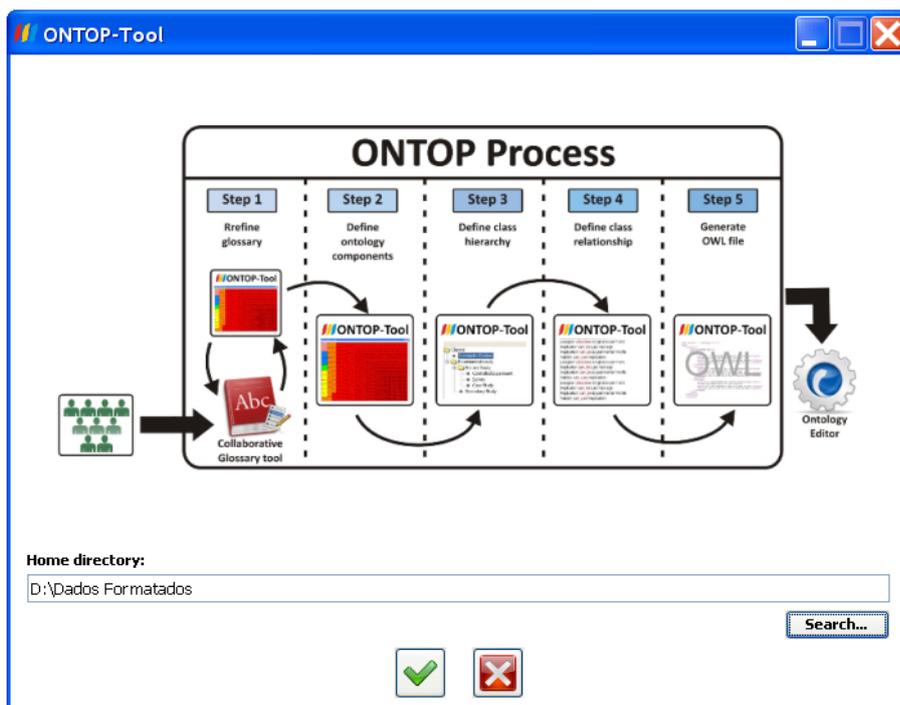


Figura 5.12. Tela inicial da ONTOP-Tool.

Nas próximas seções, são descritas as funcionalidades da ferramenta, relacionando-as com os passos do processo ONTOP, descritos na seção anterior. O uso da ferramenta, assim como do processo ONTOP, é exemplificado detalhadamente no Capítulo 6.

A Figura 5.13 apresenta a tela com as funcionalidades que permitem a execução de todos os passos do processo

Também por questão de organização, as funcionalidades providas nessa tela, que contemplam os passos 1 ao passo 5 do processo ONTOP (Figura 5.10), estão divididas em *Refine Glossary* e *Define Components*.

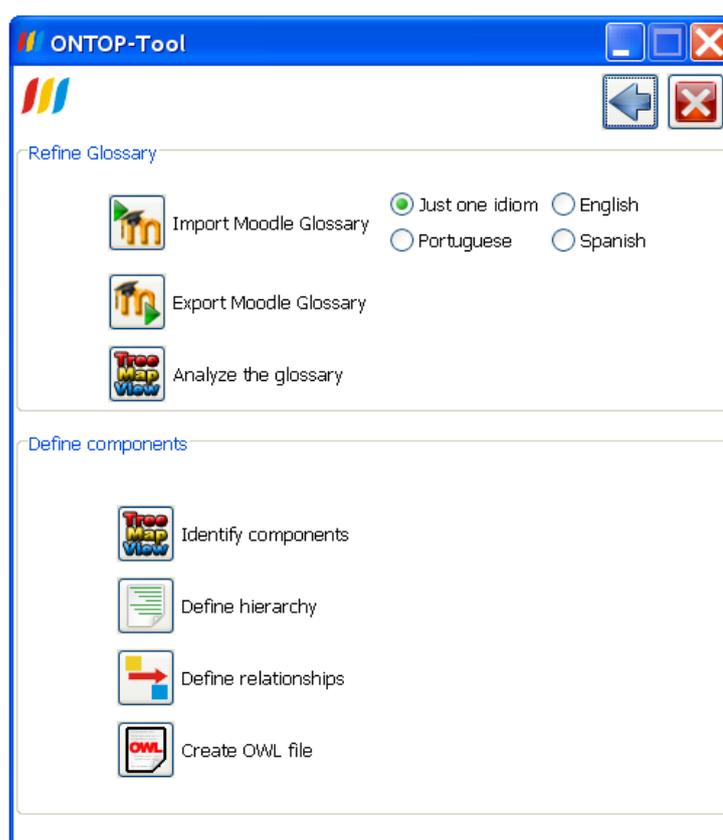


Figura 5.13. Tela com as funcionalidades da ONTOP-Tool.

As funcionalidades agrupadas em *Refine Glossary* possuem o mesmo objetivo: refinar o glossário até que este represente o domínio em questão. Essa tarefa inicia-se com a importação do glossário do Moodle para a ferramenta ONTOP-Tool. Para isso, o usuário deve exportar o glossário que está no ambiente Moodle, o que gera um arquivo XML.

Como mencionado anteriormente, a entrada dessa etapa pode ser feita de duas maneiras: ou o glossário é gerado a partir dos metadados analisados no

processo P-LIDE em que, ao final, tem-se a possibilidade de gerar uma lista de termos e importá-la no Moodle, onde os termos são definidos, ou o glossário é criado diretamente no ambiente. Em qualquer das opções o glossário deve estar no Moodle, de onde é gerado o arquivo XML, que é a entrada para o Passo 1.

O fato do glossário estar no Moodle permite que ele seja compartilhado com inúmeras pessoas que, independentemente da sua localização, podem colaborar com seu refinamento, uma vez que têm conhecimento ou são especialistas do domínio em questão e, conseqüentemente, podem contribuir para a conceitualização de uma ontologia.

O Moodle possui diversos recursos que permitem gerenciar o glossário, como por exemplo, dar níveis de acesso diferentes aos usuários e exigir que a definição de um termo, inserida pelos usuários, seja validada por alguém antes de fazer parte definitivamente do glossário. Na Figura 5.14 é apresentado um exemplo de um glossário do Moodle. É possível observar recursos de busca, de ordenação por ordem alfabética, por categoria, data de inserção e autor, e também a opção de inserir um novo item.



Figura 5.14. Tela de glossário no Moodle.

Após a importação, o usuário deve analisar e refinar o glossário. Para auxiliar nessa tarefa, a ONTOP-Tool utiliza novamente a técnica *Tree-Map*, possibilitando que o usuário visualize os termos por meio de uma metáfora visual. Para isso, basta importar o glossário do Moodle, o que gera um arquivo do tipo texto (.txt).

Esse é o arquivo que deve ser selecionado quando o usuário clicar no botão referente à funcionalidade para visualizar o glossário (botão com o rótulo “Tree-Map View”, acompanhado da descrição “Analyze the glossary”). A Figura 5.15 ilustra a tela de visualização de um glossário.

Cada retângulo da metáfora visual representa um termo, sendo que a cor e o tamanho de cada retângulo são definidos de acordo com a frequência de cada termo.

A frequência de cada termo é calculada no ato da importação, e corresponde ao número de vezes que o termo foi utilizado no conjunto de definições de todos os termos. Termos que foram utilizados mais vezes ficam evidenciados na visualização, o que permite ao usuário decidir se esse termo deve ser desconsiderado e, talvez, ser adicionado na lista de *stopwords*, ou se esse termo realmente é relevante ao domínio, sendo também um forte candidato a tornar-se um componente da ontologia.

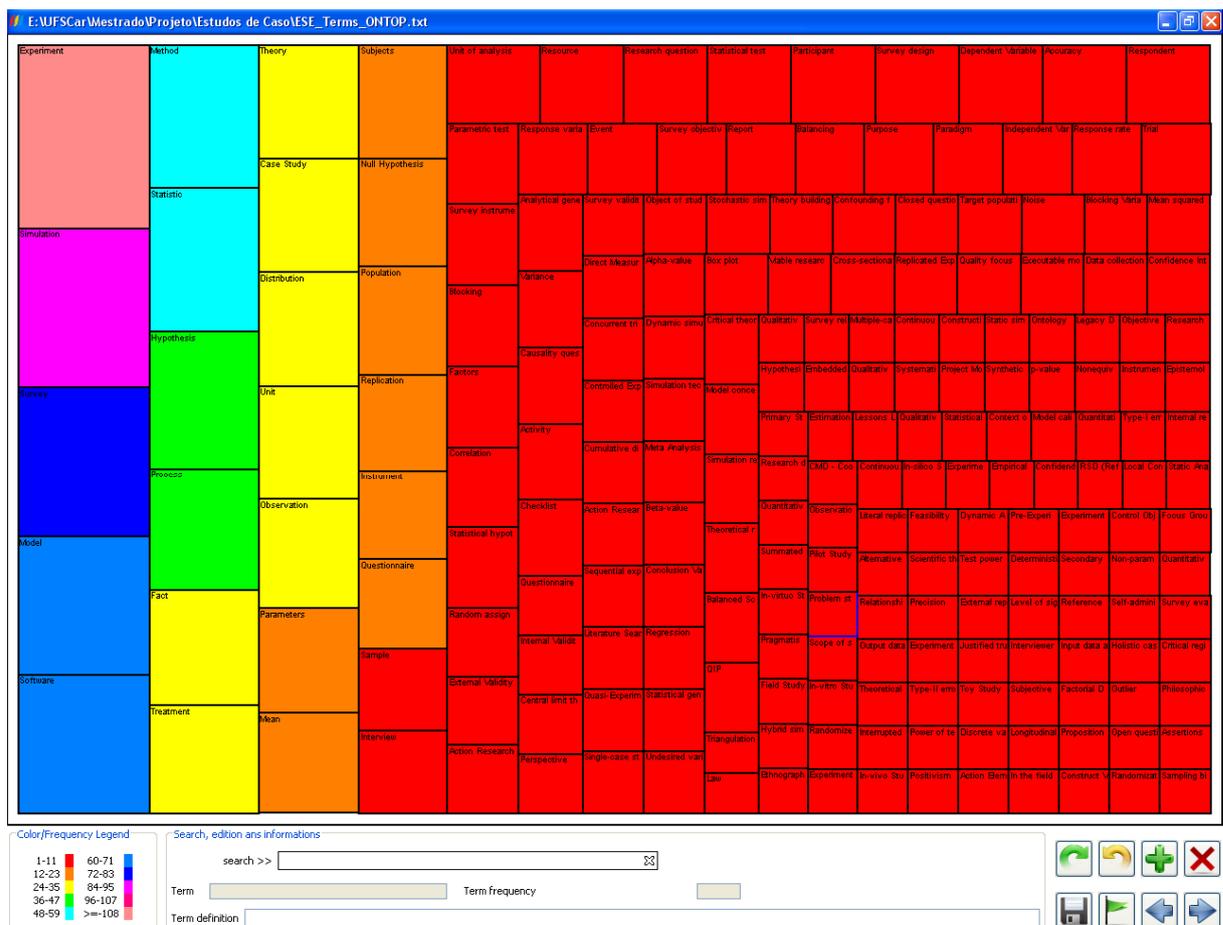


Figura 5.15. Tela da ONTOP-Tool para visualização do glossário.

Nota-se que na parte inferior da tela da Figura 5.15 é disponibilizado o recurso de busca (*search*) e informações sobre o termo. Ao clicar em um retângulo, a frequência e a descrição do termo que ele representa são disponibilizadas. O usuário pode alterar ou então inserir a descrição do termo. Termos com definição idêntica ficam agrupados, assim como acontece com as questões, no processo P-LIDE. Obviamente essa é uma situação pouco provável, mas casos em que os termos estão sem definição, essa característica da ferramenta indica ao usuário que ele deve defini-los.

Uma legenda das cores utilizadas e da faixa de frequência que cada uma representa também é disponibilizada na parte inferior da tela, no canto esquerdo, como é possível observar na Figura 5.15.

É possível que o usuário exclua ou inclua um novo termo. Nesse caso, a frequência do novo termo é calculada automaticamente pela ferramenta.

Retângulos que representam termos que foram editados ou inseridos ficam com a coloração mais clara, permitindo ao usuário identificá-los.

Na parte inferior da tela, no canto direito, são disponibilizados botões para que o usuário possa confirmar a edição do termo (botão com uma seta verde apontando para a direita), limpar a alteração feita na definição do termo (botão com uma seta amarela apontando para a esquerda), adicionar um termo (botão com um sinal “+” verde), excluir o termo selecionado (botão com a letra “X” em vermelho), salvar o glossário que está refinando para continuar a tarefa posteriormente (botão com o desenho de um disquete), finalizar o arquivo (botão com o desenho de uma bandeira verde), voltar para a tela principal (botão com uma seta azul apontando para a esquerda), ou então seguir para o próximo passo (botão com uma seta azul apontando para a direita).

É possível também que esse glossário seja exportado em formato XML compatível com o Moodle e importado no ambiente Moodle novamente, substituindo o glossário inicial ou criando um novo. Dessa forma, o usuário pode compartilhar, sempre que desejar, as novas versões do glossário. Para isso, basta clicar no botão referente a essa funcionalidade na ferramenta, que encontra-se na tela principal, selecionar o arquivo do glossário, que, assim como os questionários, é um arquivo texto separado por tabulações (.txt), e gerar o XML. Feito isso, é necessário acessar o ambiente Moodle e importar o arquivo.

Com isso, a ferramenta *ONTOP-Tool* promove uma interação entre todos os envolvidos, possibilitando a disseminação do glossário.

O usuário pode importar o glossário no Moodle e na ferramenta e exportá-lo para o Moodle quantas vezes julgar necessário. Quando a tarefa for finalizada, o usuário deve finalizar o arquivo na *ONTOP-Tool*, clicando no botão indicado, completando assim o Passo 1 e permitindo que o usuário avance para Passo 2.

Com o glossário refinado, contendo apenas termos relevantes ao domínio tratado, a ferramenta *ONTOP-Tool* permite então classificá-los como componentes da ontologia desejada.

Essa funcionalidade, assim como as próximas, está agrupada no bloco *Define components*, como pode ser observado na Figura 5.13. O objetivo dessas funcionalidades é ajudar na tarefa de conceitualização da ontologia pretendida, identificando os componentes que a irão compor. Essa tarefa inicia-se com a execução do Passo 2.2, no qual os termos são classificados.

É válido ressaltar que as funcionalidades agrupadas em *Define components* devem ser executadas sequencialmente, pois um passo depende da execução do anterior, sendo que os Passos 3 e 4 são opcionais.

Ao finalizar o glossário no Passo 1, é gerado, automaticamente, um arquivo com a extensão “ontop”, o que indica que este é um arquivo da ferramenta que será utilizado nas *Define Components*. Ao selecionar a primeira opção desse grupo de funcionalidades, que é a de identificar componentes (*Identify components*), o usuário seleciona o arquivo com a extensão “ontop”, e este é visualizado, novamente por meio da técnica de visualização *Tree-Map*, como ilustra a Figura 5.16.

Assim como na visualização inicial do glossário, cada retângulo da metáfora visual representa um termo, sendo que a cor e o tamanho de cada retângulo são definidos de acordo com a frequência de cada termo, calculada da mesma maneira do Passo 1.

Neste passo, a característica da *ONTOP-Tool* de evidenciar os termos que possuem frequência maior, chama a atenção do usuário para esses termos que, provavelmente, podem ser componentes da ontologia.

Como é possível observar na Figura 5.16, na parte inferior da tela, assim como nas outras telas, informações sobre cada termo são disponibilizadas, assim que o usuário clicar sobre cada retângulo. A legenda das cores utilizadas e da faixa de frequência que cada uma representa também está presente nessa tela.

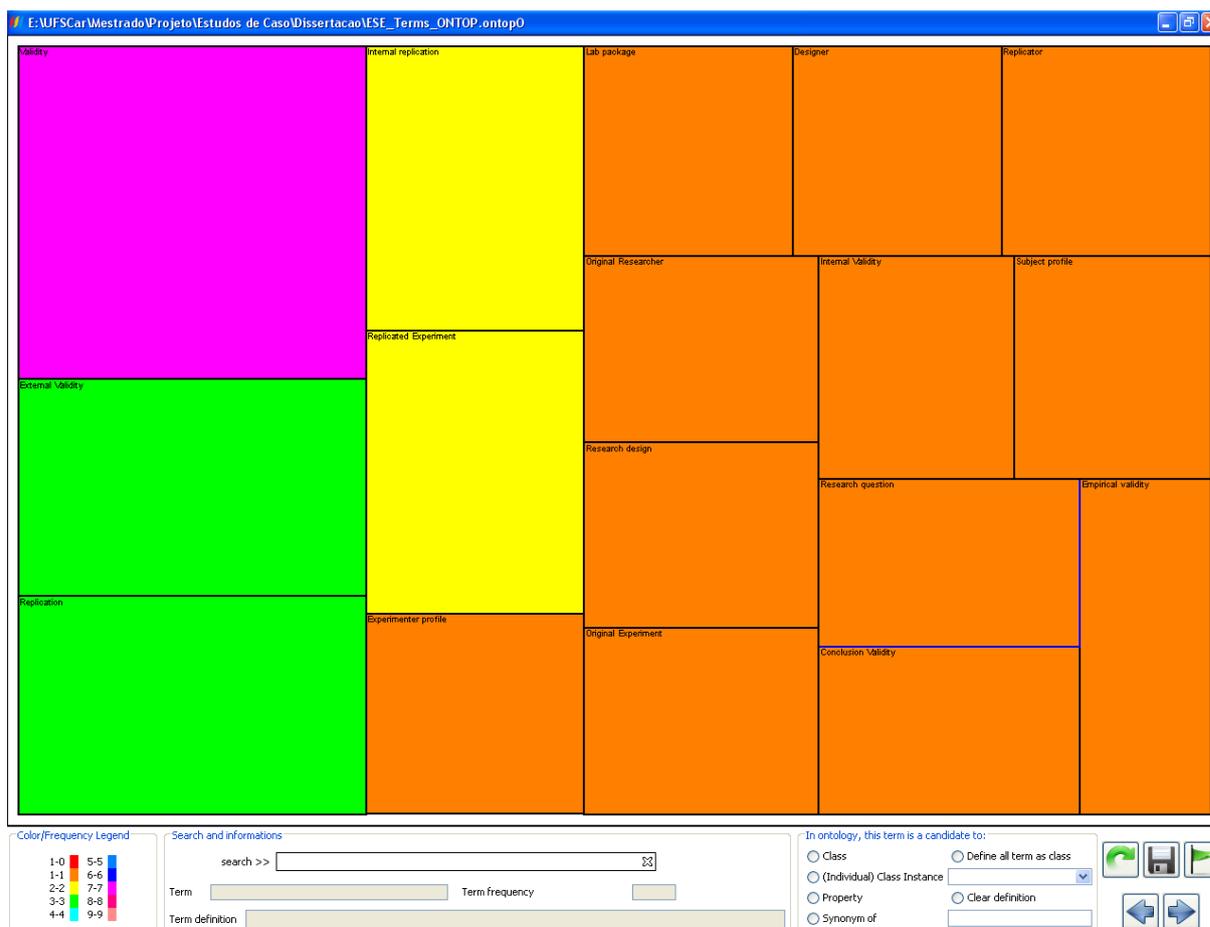


Figura 5.16. Tela da ONTOP-Tool para classificação dos termos.

Nesse passo, não é permitido que o usuário altere a descrição dos termos. O usuário deve apenas classificá-lo como:

- Classe;
- Indivíduo (instância de uma classe), e nesse caso o usuário deve indicar a classe da qual o termo é uma instância;
- Propriedade (relação);
- Sinônimo de outro termo e, nesse caso, o termo do qual ele é sinônimo também deve ser indicado.

A ferramenta ainda fornece a opção para que o usuário desfaça uma classificação feita ou então defina todos os componentes da tela como classe, caso todos os termos sejam considerados classes da ontologia.

O retângulo que representa um termo já definido, assim como nas outras metáforas visuais usadas na ferramenta, fica com a coloração mais clara.

Na parte inferior da tela, no canto direito, são disponibilizados botões para que o usuário possa confirmar a classificação (botão com uma seta verde apontando para a direita), salvar o arquivo para continuar a classificação depois (botão com o desenho de um disquete), finalizar o arquivo (botão com o desenho de uma bandeira verde), voltar para a tela principal (botão com uma seta azul apontando para a esquerda), ou então seguir para o próximo passo (botão com uma seta azul apontando para a direita).

Ao finalizar esse passo, o usuário finaliza o arquivo por meio do botão indicado, o que gera um arquivo com a extensão “ontopc”. Dessa forma, contempla-se o Passo 2 do processo ONTOP, como pode ser observado na Figura 5.10, o que permite que o usuário avance para o Passo 2.3 do processo.

O Passo 3 do processo ONTOP é apoiado pela funcionalidade representada na Figura 5.17. Por meio dessa tela, o usuário pode organizar hierarquicamente as classes definidas no passo anterior. Para isso, o usuário clica no botão referente a essa funcionalidade e seleciona o arquivo com a extensão “ontopc”. A ferramenta ONTOP-Tool fornece uma interface *drag-and-drop* para que a hierarquização seja realizada.

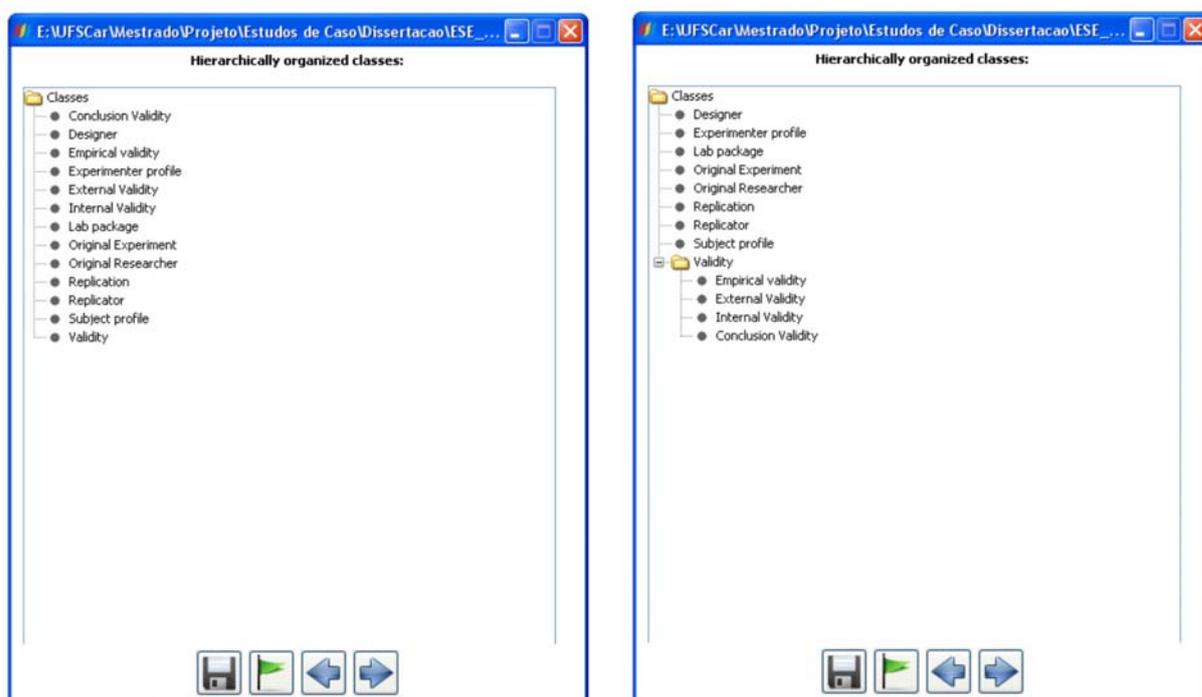


Figura 5.17. Tela da ONTOP-Tool para hierarquização das classes.

Inicialmente, as classes definidas no Passo 2 são mostradas sequencialmente ao usuário, em ordem alfabética. Executando a ação de *drag-and-drop*, o usuário organiza as classes criando a hierarquia adequada.

Na parte inferior da tela, são disponibilizados botões para que o usuário possa salvar o arquivo e continuar a tarefa em outro momento (botão com o desenho de um disquete), finalizar o arquivo (botão com o desenho de uma bandeira verde), seguir para o próximo passo (botão com uma seta azul apontando para a direita) ou então voltar para a tela principal (botão com uma seta azul apontando para a esquerda).

A execução do Passo 3 é opcional, pois o usuário pode preferir executar essa tarefa em uma ferramenta específica para criação e edição de ontologias, como as mencionadas no Capítulo 4, ou pode desejar usar a ferramenta apenas para abstrair os componentes da ontologia com apoio da visualização.

Com a hierarquização feita, basta o usuário finalizar o arquivo, clicando no botão adequado, gerando um arquivo com a extensão “ontoph”, completando assim o Passo 3 do processo ONTOP e permitindo que o usuário avance para o Passo 4.

No Passo 4 o usuário pode definir alguns relacionamentos *Domain-Range* (descritos no Capítulo 4), usando as classes e as propriedades definidas no Passo 2. Para executar essa tarefa, o usuário clica no botão referente a essa funcionalidade e seleciona o arquivo com a extensão “ontoph”, quando houver a hierarquização, ou “ontopc”, quando a execução do Passo 2 não for realizada. A Figura 5.18 ilustra a tela que implementa essa funcionalidade.

Nota-se que as classes definidas no Passo 2 são disponibilizadas em duas listas, uma do lado esquerdo e outra do lado direito da tela. Entre essas listas, no centro da parte superior da tela, são listadas as propriedades, também definidas no mesmo passo.

Para criar um relacionamento *Domain-Range* (domínio-escopo), o usuário deve selecionar a classe *Domain*, na lista do lado esquerdo, selecionar a propriedade na lista do centro e depois selecionar a classe *Range* na lista de classes do lado direito e confirmar, clicando no botão indicado (botão com o desenho de uma seta verde apontando para a direita). Caso a propriedade desejada não exista, é possível inseri-la nesse momento.

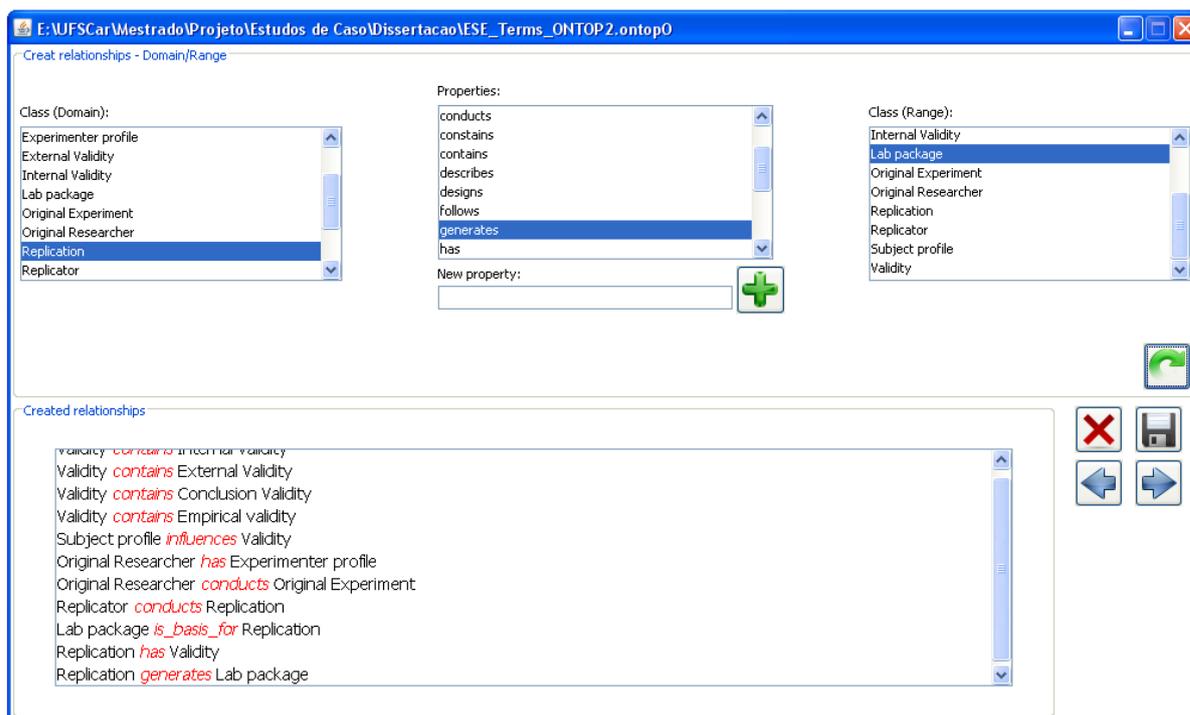


Figura 5.18. Tela da ONTOP-Tool para definição de relacionamentos.

Conforme o usuário define os relacionamentos, eles são listados na parte inferior da tela, do lado esquerdo.

Na parte inferior, do lado direito encontram-se os botões que permitem ao usuário excluir um relacionamento criado (botão com a letra “X” em vermelho), salvar o arquivo (botão com o desenho de um disquete), voltar para a tela principal (botão com uma seta azul apontando para a esquerda), ou então seguir para a execução do Passo 5, que é a geração do arquivo OWL (botão com uma seta azul apontando para a direita). Nessa funcionalidade não é necessário finalizar um arquivo, pois a primeira vez que o usuário salvar o arquivo, este já será salvo com a extensão “ontopr”.

Assim como a funcionalidade do Passo 3, a do Passo 4 também é opcional, pois o usuário pode não definir domínios e seus respectivos escopos (*Domain-Range*) por meio da ONTOP-Tool.

Com o Passo 4 do processo ONTOP terminado, o usuário pode finalizar o processo, executando o Passo 5, que permite gerar o arquivo OWL. Para isso, ou o usuário escolhe a opção disponível na tela de execução do Passo 4 (botão com uma seta apontando para a direita), ou então clica no botão referente a essa funcionalidade (botão com o rótulo “OWL”) na tela das funcionalidades, como pode ser observado na Figura 5.13.

Escolhendo uma das duas opções, o usuário seleciona o arquivo com a extensão “ontopr”, caso tenha realizado a hierarquização, ou o arquivo com a extensão “ontopc” e, automaticamente, a *ONTOP-Tool* gera o arquivo OWL contendo todas as informações definidas com o apoio da ferramenta.

O arquivo com a extensão “ontopr”, gerado no Passo 4, é usado para gerar o arquivo OWL, mas não precisa ser selecionado. Basta que ele esteja no mesmo diretório que o arquivo selecionado (ontopr ou ontopc) e a *ONTOP-Tool* o encontra.

Com a geração do arquivo OWL, essa pré-conceitualização da ontologia pode ser importada em uma ferramenta específica para criação e edição de ontologias, como a Protégé-2000 e outras, citadas no Capítulo 4.

Ao importar o arquivo OWL nessas ferramentas, os engenheiros de ontologia podem se concentrar apenas na tarefa de implementação dos axiomas e finalização da ontologia, pois as classes, propriedades e indivíduos já estarão instanciados na ferramenta. Não deve ser descartada a possibilidade de, em alguns casos, ser necessário criar, na ferramenta selecionada, uma nova classe, propriedade ou instância. No entanto, essa tarefa é visivelmente minimizada quando se usa a *ONTOP-Tool*.

5.6 Considerações Finais

Nesse capítulo foram apresentados os processos P-LIDE e ONTOP, criados para dar suporte às dificuldades enfrentados pelo INEP com relação aos dados de suas avaliações educacionais, como é relatado no Capítulo 1 e na Seção 5.1.

O processo P-LIDE permite a geração da LIDE, que é uma representação comum e organizada de todas as questões que formam os diversos questionários utilizados pelo INEP.

Esse processo prevê a solução de um problema prático, tendo em vista o grande volume de dados produzido por meio das avaliações e que, atualmente, se encontram armazenados sem padronização, impedindo análises mais complexas. A LIDE, gerada no P-LIDE, é a base para que a plataforma Web-PIDE (Capítulo 2) seja construída.

Durante a análise e padronização sintática necessária para a criação da LIDE,

o que é feito com a aplicação do P-LIDE e da *SEV-Tool*, notou-se a necessidade de uma ontologia para o domínio das avaliações do INEP. Essa ontologia pode padronizar os termos utilizados nas questões dos questionários que compõem as avaliações e pode formalizar o domínio, uma vez que ele é utilizado em diversas pesquisas interdisciplinares.

Em consequência, foi identificada na literatura a lacuna existente em relação a processos bem definidos e ferramentas computacionais que apoiem a fase de conceitualização de ontologias.

O processo ONTOP e a *ONTOP-Tool*, desenvolvidos para apoiar a conceitualização de ontologias, contribuem para a disciplina de engenharia de ontologia e por isso, não foram desenvolvidos especificamente para o domínio de avaliação educacionais do INEP.

Executar a fase de conceitualização de ontologias de uma maneira organizada contribui para a documentação e manutenção do artefato e, conseqüentemente, para sua vida útil.

A elaboração de um glossário de maneira colaborativa auxilia a disseminação do conhecimento sobre o domínio. Aliando o glossário ao recurso de visualização fornecido pela *ONTOP-Tool*, é possível facilitar a tarefa de identificar os componentes que irão compor a ontologia desejada.

Como observado no decorrer deste capítulo, o processo P-LIDE, que recebe apoio computacional da *SEV-Tool*, permite a análise e padronização sintática dos dados oriundos das avaliações educacionais realizadas anualmente pelo INEP.

Essa análise facilita o entendimento da estrutura dos questionários e a padronização sintática permite a geração da LIDE, artefato base para a construção da plataforma Web-PIDE, citada no Capítulo 2.

A *SEV-Tool* provê os recursos não encontrados em outra ferramenta de visualização: busca e edição, permitindo que por meio deles, o usuário consiga padronizar sintaticamente os dados, tornando os que são semanticamente iguais, sintaticamente idênticos.

O uso da visualização na ferramenta facilita a tarefa de padronização e é usada para direcionar a análise durante a atividade. A visualização conjunta de todos os questionários equivalentes permite o tratamento de uma grande quantidade de questões, tornando a padronização mais efetiva.

O processo ONTOP, que recebe o apoio computacional da *ONTOP-Tool*,

além de prover funcionalidades específicas para o projeto Web-PIDE, objetiva ser um apoio para a fase de conceitualização de ontologias, utilizando visualização.

Para apoiar a conceitualização, o processo permite a entrada de dados por meio de duas situações diferentes: quando se tem um conjunto de metadados (atualmente, os metadados do INEP) e quando se tem um glossário de termos caracterizando o domínio para o qual se pretende criar a ontologia. Independentemente do modo de entrada dos dados, a criação de um glossário é sempre necessária para permitir que a definição dos termos seja a mais fiel ao domínio em questão.

No caso do ONTOP, é utilizado o recurso de glossário do Moodle, o que possibilita uma grande interação entre os especialistas do domínio, pois o conhecimento representado no glossário é facilmente compartilhado e validado por várias pessoas.

O uso da visualização na ferramenta *ONTOP-Tool* provê facilidade para que todos os termos sejam visualizados conjuntamente, com a indicação da frequência com que aparecem no conjunto de dados. Essa característica permite que a atenção do usuário seja voltada primeiramente para os termos com maior utilização, sendo esses, fortes candidatos a se tornarem componentes da ontologia.

Ao gerar um arquivo OWL automaticamente pela *ONTOP-Tool* as próximas fases da construção da ontologia podem ser imediatamente iniciadas, poupando trabalho de criação e instanciação de todas as suas classes e relacionamentos.

Na sua versão atual, a *ONTOP-Tool* não realiza processamentos linguísticos nos dados de entrada do processo. Para auxiliar o usuário a abstrair os componentes da ontologia, é utilizada a metáfora visual, criada por meio da técnica *Tree-Map*. O cálculo da frequência com que os termos são utilizados no glossário aliado à representação visual permite que o usuário identifique rapidamente os termos mais utilizados no conjunto de dados visualizado.

Este capítulo apresentou os processos P-LIDE e ONTOP, resultados desta pesquisa, e as ferramentas que o apoiam, *SEV-Tool* e *ONTOP-Tool*.

A execução dos processos e dos respectivos passos que os compõem foram detalhados, assim como as funcionalidades das ferramentas. Exemplos da aplicação do processo serão apresentados no capítulo seguinte.

Capítulo 6

APLICAÇÃO DOS PROCESSOS P-LIDE E ONTOP

Este capítulo apresenta a aplicação dos processos P-LIDE e ONTOP com o apoio das respectivas ferramentas, SEV-Tool e ONTOP-Tool. O P-LIDE é exemplificado com dados das avaliações educacionais do INEP e o ONTOP, com um glossário de Engenharia de Software Experimental.

6.1 Considerações Iniciais

Como mencionado no Capítulo 5, a criação da LIDE foi a motivação inicial para implementação da ferramenta *SEV-Tool* e, para essa tarefa, havia a necessidade de analisar as questões dos diversos questionários usados pelo INEP no processo de avaliação do ensino, com o objetivo de identificar os anos e as provas em que cada questão foi utilizada. A visualização foi o recurso facilitador escolhido nesta pesquisa para realizar essa tarefa.

Ao iniciar a manipulação dos arquivos disponibilizados pelo INEP a fim de entender como os arquivos são estruturados e como deveria ser feita a análise, além do grande volume de dados para ser processado, alguns problemas com esses dados foram detectados. Esses problemas fizeram com que os esforços para a construção da LIDE nesta pesquisa se concentrassem somente na avaliação SAEB que, a pedido do INEP, deveria ser a avaliação inicialmente utilizada no projeto Web-PIDE.

Para executar o processo ONTOP, é necessária a interação com membros do

INEP, responsáveis por gerir os questionários. Como essa tarefa é lenta e envolve alguns trâmites, embora a intenção fosse mostrar a aplicação dos dois processos com base no mesmo contexto, isso não foi possível em decorrência das barreiras encontradas junto aos membros do INEP. Assim, decidiu-se exemplificar o ONTOP e o uso da *ONTOP-Tool* com um glossário sobre Engenharia de Software Experimental, construído com a participação de pesquisadores da área, sob a coordenação da orientadora desta pesquisa.

Neste capítulo é apresentada a execução do processo P-LIDE com o apoio da *SEV-Tool* e do processo ONTOP com apoio da *ONTOP-Tool*, exemplificando a aplicação dos processos e a execução das ferramentas, apresentados nos capítulos anteriores. A Seção 6.2 apresenta a preparação dos dados do SAEB para serem utilizados pela *SEV-Tool*. Na Seção 6.3 é apresentada a execução do P-LIDE, que prevê a análise e a padronização sintática dos questionários do INEP e a geração da LIDE. Na Seção 6.4 é apresentada a execução do ONTOP, que prevê passos para auxiliar a conceitualização de uma ontologia com base em um glossário. As considerações finais são feitas na Seção 6.5.

6.2 Processo P-LIDE: um exemplo com os dados do INEP

Concentrando esforços nos dados do SAEB, de acordo com a solicitação do INEP, nesta seção exemplifica-se a execução da Etapa1 do processo ONTOP para a criação da LIDE_{SAEB}. Ressalta-se que, para as demais avaliações o mesmo procedimento adotado com as bases do SAEB deve ser adotado com as outras bases para que as respectivas LIDEs sejam geradas. Como mencionado no Capítulo 5, os dados disponibilizados pelo INEP precisam de um tratamento inicial para serem carregados na *ONTOP-Tool*, preparação esta descrita na seção a seguir.

6.2.1 Passo 1: Preparar os dados do SAEB para serem importados na *SEV-Tool*

Como apresentado no Capítulo 2, embora o objetivo inicial do projeto Web-PIDE fosse tratar as bases de dados educacionais de todas as avaliações do INEP,

usando os dados do ensino superior como dados iniciais, logo no início do projeto o INEP informou a necessidade de iniciar os estudos com as bases de dados do SAEB, pois é nesse contexto em que as análises são mais urgentes.

Com os microdados do SAEB disponibilizados pelo INEP, a equipe do projeto criou o banco de dados `webpide_saeb` no SGBD (Sistema de Gerenciamento de Banco de Dados) PostgreSQL para realizar o armazenamento e centralização dos microdados. Inicialmente ocorreu a criação dos dicionários de variáveis para mapear os atributos nos microdados e, em seguida, estes foram inseridos na base `webpide_saeb`.

Para a criação do dicionário de variáveis, a equipe da Universidade Federal de São Carlos (UFSCar) desenvolveu uma aplicação intitulada Carga de Microdados WebPIDE (CMWP), com o intuito de permitir que os microdados, em arquivos texto, sejam carregados no SGBD PostgreSQL. A definição dessas tabelas são feitas de acordo com a documentação presente nos arquivos “leia-me”.PDF citados no Capítulo 2.

No entanto, para a definição da LIDE e das LIDEs específicas de cada avaliação, é necessário analisar os metadados das bases de dados.

Sobre o SAEB, o INEP disponibilizou doze questionários diferentes, com dados das suas aplicações de 1995 a 2003, sendo que para cada questionário há variações das séries aos quais eles se destinam. Na Tabela 7.1 é possível visualizar todos os questionários do SAEB utilizados nesta pesquisa, totalizando sessenta e seis arquivos.

O maior problema enfrentado inicialmente foi que as informações contidas nos três tipos de documentos disponibilizados pelo INEP (PDF, ASCII e SAS), diferentemente do que se esperava, não são idênticas. Um exemplo pode ser obtido ao analisar o arquivo PDF e o arquivo SAS do questionário do SAEB “DIRETOR_03”, no qual o atributo “D154_2” não é apresentado no PDF, mas está presente no arquivo SAS, no qual possui a descrição “Agressão verbal a professores - professores”. O mesmo acontece com outros atributos.

Essa característica exigiu que a equipe do projeto realizasse uma comparação manual dos arquivos para detectar as diferenças e possibilitar a decisão de qual dos formatos de arquivo retrata fielmente a estrutura dos microdados. No Apêndice A é possível encontrar o relatório feito pela equipe do projeto Web-PIDE para relatar os problemas encontrados entre os diferentes arquivos disponibilizados

pelo INEP.

Após a análise dos arquivos, decidiu-se utilizar os arquivos em formato SAS, por este ser um formato de fácil manipulação.

Tabela 6.1. Questionários do SAEB utilizados na pesquisa.

Questionário	1995	1997	1999	2001	2003
Biologia		BIOLOGIA_03ANO	BIOLOGIA_03ANO		
Ciências		CIENCIAS_04SERIE CIENCIAS_08SERIE	CIENCIAS_04SERIE CIENCIAS_08SERIE		
Diretor	DIRETOR_95	DIRETOR_97	DIRETOR_99	DIRETOR_01	DIRETOR_03
Docentes	DOCENTES_95	DOCENTES_97	DOCENTES_99	DOCENTES_01	DOCENTES_03
Escola	ESCOLA_95	ESCOLA_97	ESCOLA_99	ESCOLA_01	ESCOLA_03
Física		FISICA_03ANO	FISICA_03ANO		
Geografia			GEOGRAFIA_04SERIE GEOGRAFIA_08SERIE GEOGRAFIA_03ANO		
História			HISTORIA_04SERIE HISTORIA_08SERIE HISTORIA_03ANO		
Matemática	MATEMATICA_04SERIE MATEMATICA_08SERIE MATEMATICA_03ANO	MATEMATICA_04SERIE MATEMATICA_08SERIE MATEMATICA_03ANO	MATEMATICA_04SERIE MATEMATICA_08SERIE MATEMATICA_03ANO	MATEMATICA_04SERIE MATEMATICA_08SERIE MATEMATICA_03ANO	MATEMATICA_04SERIE MATEMATICA_08SERIE MATEMATICA_03ANO
Português	PORTUGUES_04SERIE PORTUGUES_08SERIE PORTUGUES_03ANO	PORTUGUES_04SERIE PORTUGUES_08SERIE PORTUGUES_03ANO	PORTUGUES_04SERIE PORTUGUES_08SERIE PORTUGUES_03ANO	PORTUGUES_04SERIE PORTUGUES_08SERIE PORTUGUES_03ANO	PORTUGUES_04SERIE PORTUGUES_08SERIE PORTUGUES_03ANO
Química		QUIMICA_03ANO	QUIMICA_03ANO		
Turma			TURMA_04SERIE TURMA_08SERIE TURMA_03ANO	TURMA_01	TURMA_03

Para que os arquivos SAS fossem utilizados na *SEV-Tool*, foi necessário reunir todos os arquivos correspondentes a um questionário (todos os arquivos do questionário de português, por exemplo) em um só arquivo, no formato texto, separados por tabulações (.txt). Ao executar essa tarefa, notou-se que nem todos os arquivos SAS seguiam a mesma tabulação. Em alguns casos, a posição do nome do atributo era igual, por exemplo, ao tipo do atributo no banco de dados.

Para resolver esse problema, todos os arquivos SAS foram reunidos, manualmente, em um arquivo texto para que fosse feita a formatação da tabulação adequada em todos eles ao mesmo tempo.

Detectou-se que em alguns arquivos SAS a descrição do atributo era nula. Nesses casos foram necessárias consultas aos arquivos “leia-me.pdf” para inserir no arquivo SAS a descrição correta do atributo.

Ao unificar os arquivos SAS, sentiu-se a necessidade de adicionar informações sobre o ano em que o questionário foi aplicado e também o nome do questionário, pois essas informações possibilitam identificar a série em que o questionário foi aplicado. Assim, mesmo quando todos os arquivos SAS estão no mesmo arquivo texto, atributos com nome idêntico em mais de um questionário são facilmente distinguidos.

Por fim, os sessenta e seis questionários do SAEB foram transformados em doze arquivos de metadados (pois os dados contidos em cada linha do arquivo texto são informações sobre os dados do banco de dados nela representado), sendo que esses foram a base para a criação da LIDE.

6.2.2 Passo 2: Gerar a LIDE

Os doze arquivos contendo os sessenta e seis questionários do SAEB foram visualizados na *ONTOP-Tool*, analisados e padronizados sintaticamente.

Na Figura 6.1 é apresentada a tela da visualização inicial do arquivo de metadados dos questionários de português, que é formado por quinze questionários. Nessa figura, cada retângulo colorido corresponde a uma questão de um questionário. Observando o lado esquerdo da tela, é possível verificar que apenas quatro questões eram sintaticamente iguais e estavam presentes, em todos os questionários, pois existem 15 ocorrências agrupadas das mesmas, o que corresponde as 15 diferentes provas da disciplina de “português”, já realizadas nas avaliações do SAEB. No entanto, observando o lado direito da Figura 6.1, principalmente no canto inferior, pode-se ver que um grande número de questões são sintaticamente iguais apenas em três ou menos questionários, pois esse é o número de retângulos que se encontram agrupados. Nota-se também que várias questões foram utilizadas somente em um questionário de um determinado ano, visto que existem vários retângulos que estão isolados (exatamente no canto inferior direito).

Assim, a partir desse ponto, com base nessa situação encontrada, foi preciso analisar cada uma das questões para verificar se ela não era semanticamente igual a nenhuma outra, mesmo que com sintaxe diferente. Em outras palavras, foi preciso verificar se a questão estava escrita de forma diferente, mas tinha a intenção de fazer o mesmo questionamento.

Clicando em cada retângulo, que representa uma questão, é possível verificar a semelhança semântica entre as questões.

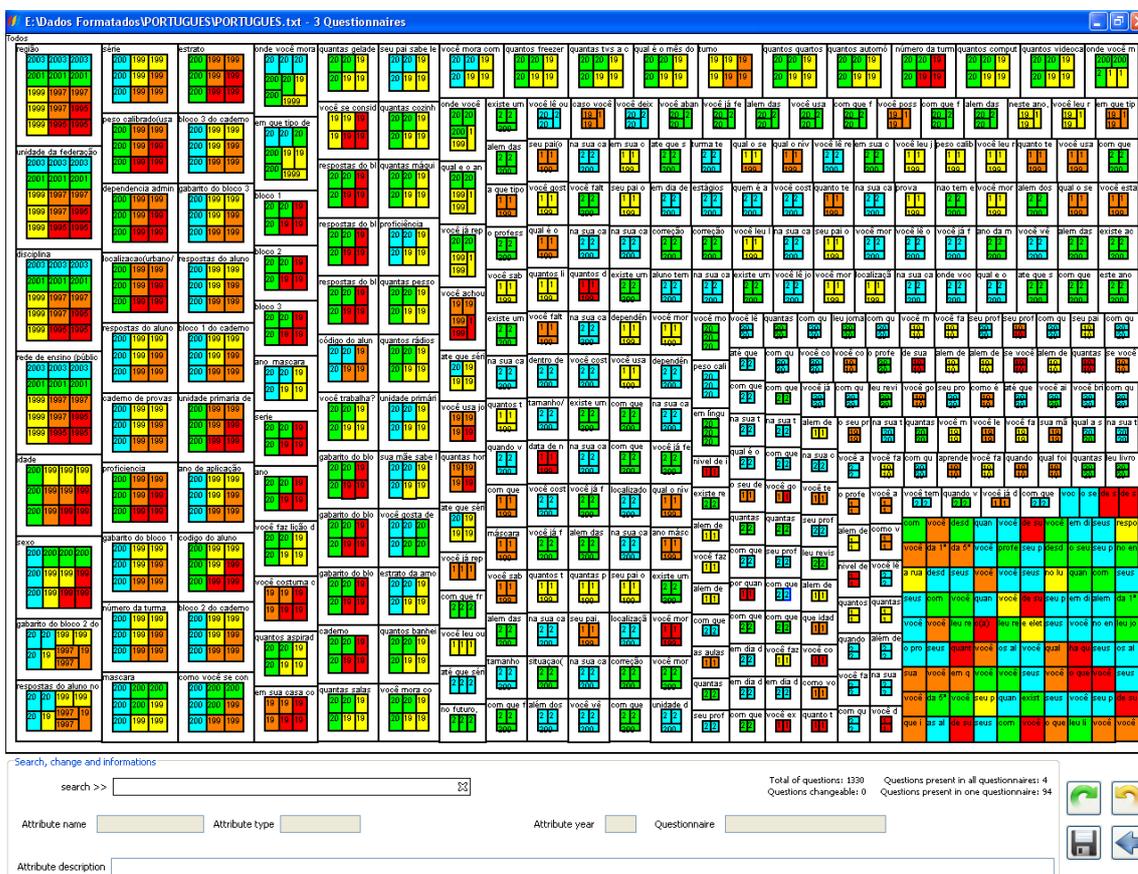


Figura 6.1. Tela de visualização dos questionários de português do SAEB.

Na Figura 6.2 é possível observar o primeiro passo da padronização sintática. Para essa tarefa, o seguinte roteiro foi seguido:

- 1) Clicar em uma questão isolada, que permanece destacada na cor azul até que o usuário clique em outra questão, como pode ser observado na Figura 6.2 no canto inferior direito, destacado com um triângulo. No exemplo: (“*you conclude the fundamental(1st grade) in the suppletive?*”);
- 2) Identificar uma palavra-chave na questão. No exemplo: (“*suppletive*”);
- 3) Fazer uma busca de questões que são compostas pela palavra-chave utilizando o recurso de busca da ONTOP-Tool;
- 4) Com as questões que são compostas pela palavra-chave e que ficam destacadas com a cor cinza (marcadas com círculos na Figura 6.2),

analisar se elas são semanticamente iguais e definir a sintaxe mais adequada para padronizá-las;

No exemplo:

- “*you concluded the 1st degree in supplementary education?*”, descrição do atributo “A031_015” do questionário PORTUGUES_03ANO e também do atributo “A081_015” do questionário PORTUGUES_08SERIE, ambos de 1995;
- “*you concluded the fundamental (old 1st degree) in supplementary education?*”, descrição do atributo “A112L007” do questionário PORTUGUES_03ANO de 2003;
- “*you concluded the fundamental(1st degree) in supplementary education?*”, descrição do atributo “A032_009” do questionário PORTUGUES_03ANO de 1999;
- “*you concluded the fundamental in supplementary education?*”, descrição do atributo “A112L036” do questionário PORTUGUES_03ANO de 2001;

5) Com a sintaxe adequada escolhida, todas as questões destacadas pela busca são alteradas, uma a uma, por meio do recurso de edição da ONTOP-Tool.

- No exemplo, a sintaxe escolhida foi: “*you concluded the fundamental (old 1st degree) in supplementary education?*”;

6) Assim que cada questão é editada, a visualização é automaticamente modificada para que as questões sintaticamente idênticas sejam agrupadas. Na Figura 6.3 é apresentada a tela de visualização após a edição das questões do exemplo, sendo que essas estão destacadas por um círculo. É possível notar que a cor dos retângulos que representam uma questão editada está mais clara, indicando que as mesmas já foram analisadas;

7) Finalizada a edição de todas as questões destacadas, uma nova questão isolada é selecionada e volta-se ao passo descrito no número 1.

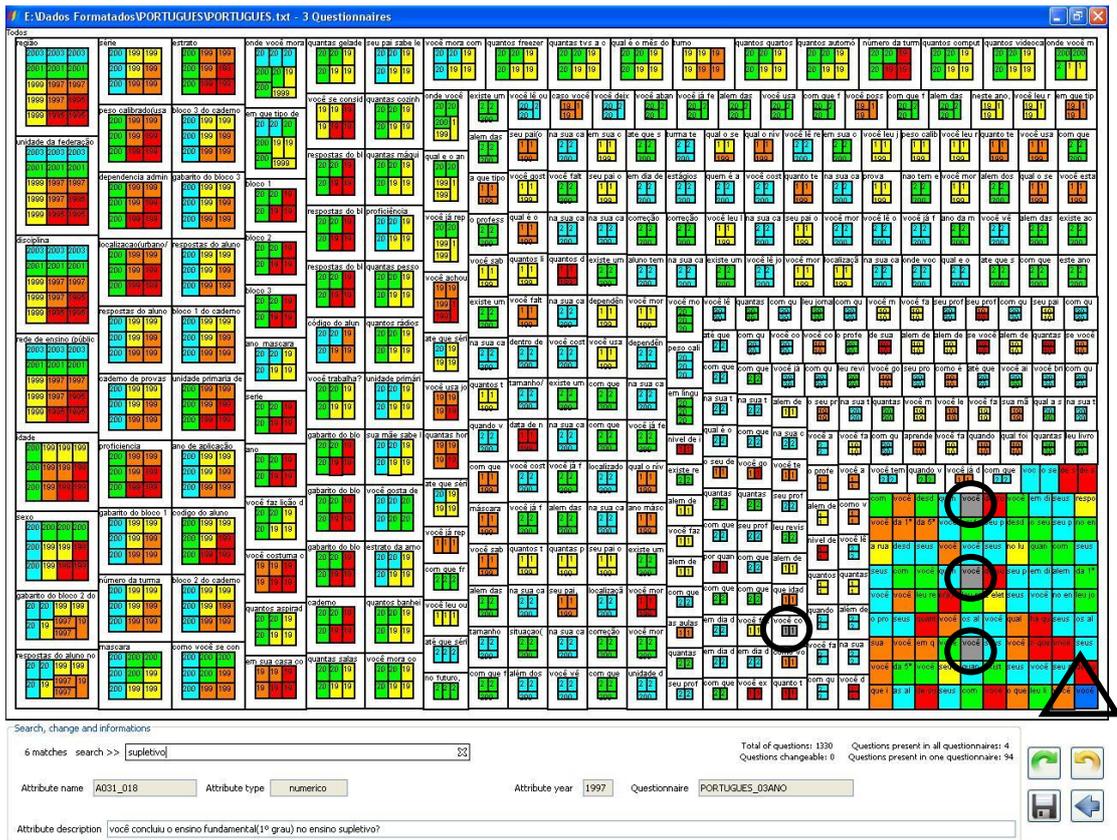


Figura 6.2. Tela da ONTOP-Tool/ no início da padronização sintática.

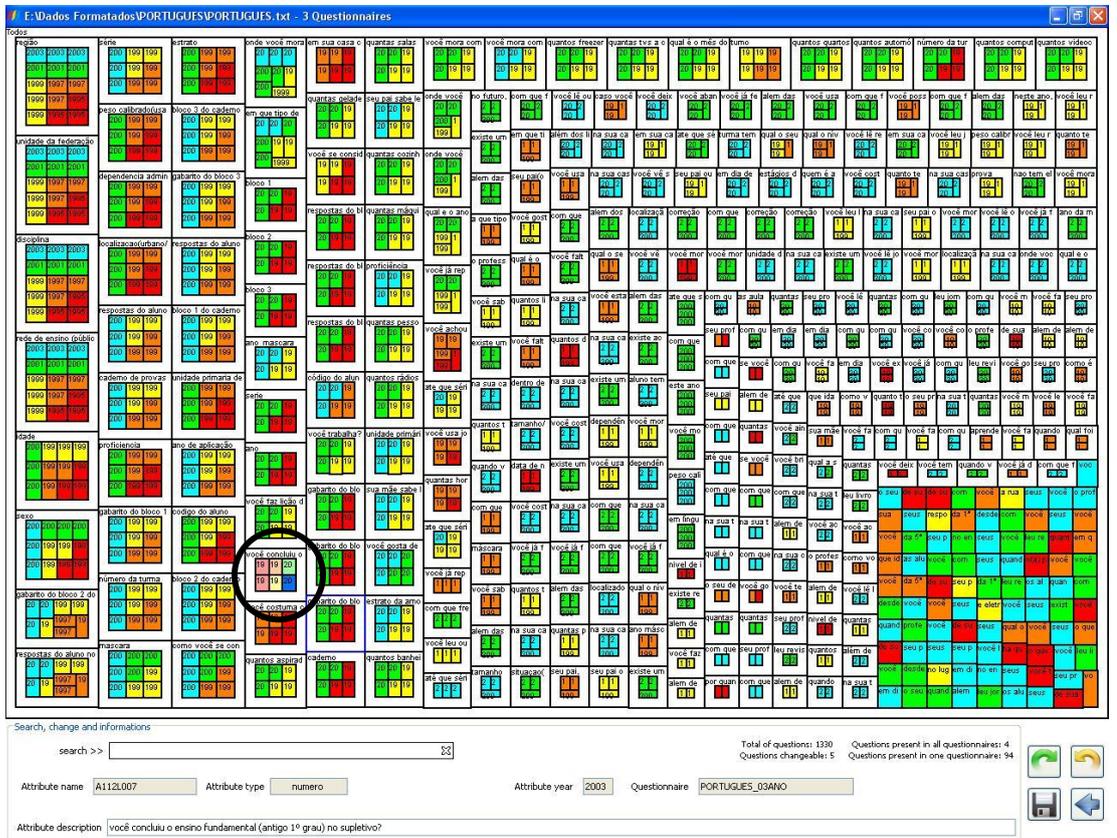


Figura 6.3. Tela da ONTOP-Tool/ após a padronização de cinco questões.

Quando as questões isoladas forem todas padronizadas, a seleção passou a ser de questões com apenas duas utilizações nos questionários e assim por diante, até que todas as questões sejam analisadas e padronizadas, quando possível.

Na Figura 6.4 é apresentada a tela da visualização após todas as modificações serem realizadas.

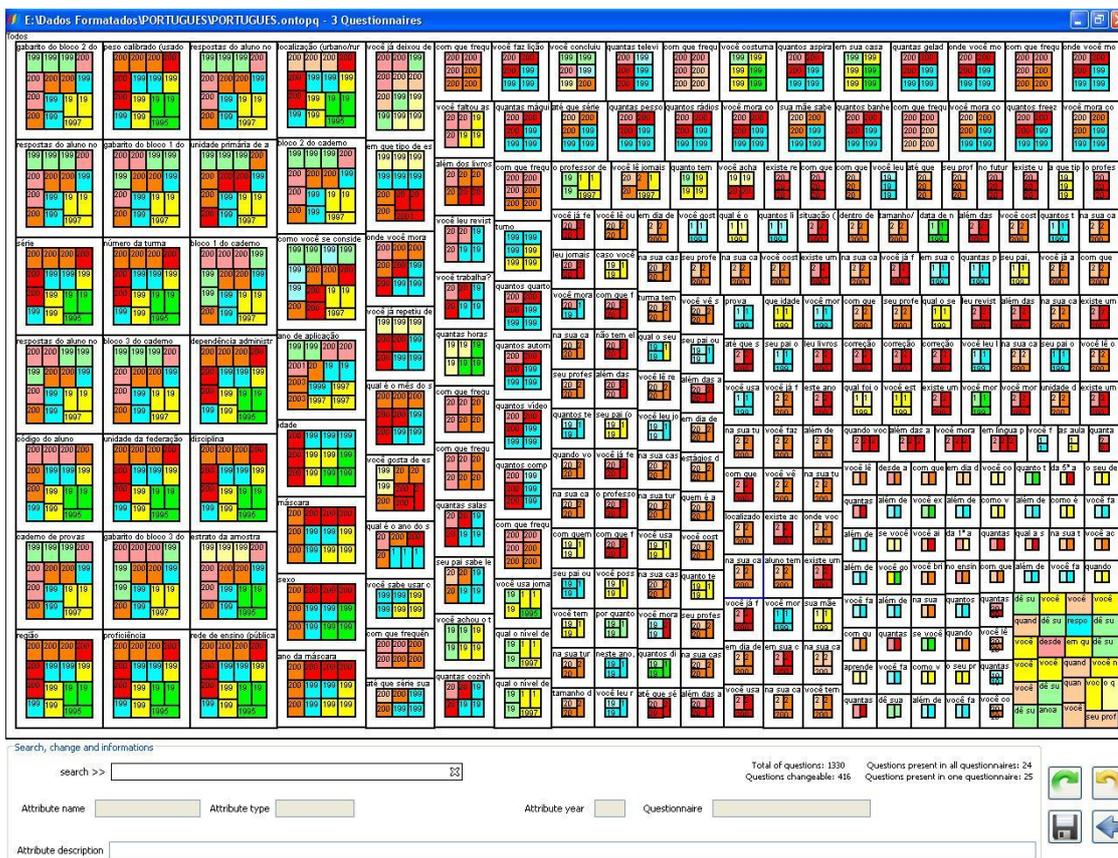


Figura 6.4. Tela da ONTOP-Tool após a padronização de todas as questões.

Em alguns casos a padronização das questões não foi possível, pois apesar da mínima diferença, a modificação poderia ter consequências indesejadas, pois, em alguns casos, embora muito parecidas sintaticamente, a intenção da questão pode ser diferente. Um exemplo pode ser observado na questão: “*você leu revistas em quadrinhos neste ano?*”, do questionário PORTUGUES_03ANO de 1999. A mesma questão está presente, com a mesma sintaxe, nos questionários PORTUGUES_04SERIE e PORTUGUES_08SERIE do mesmo ano. Já no questionário PORTUGUES_03ANO de 2001, essa questão também está presente, embora com a sintaxe levemente diferente: “*leu revistas em quadrinhos este ano*”. Nesse caso, não há dúvidas de que elas são a mesma questão e que adotar uma única sintaxe para elas não trará problemas. Entretanto, nos três questionários de

2003, PORTUGUES_04SERIE, PORTUGUES_08SERIE e PORTUGUES_03ANO, a questão aparece redigida da seguinte forma: “*você lê revistas em quadrinhos?*”. Nesse caso a sintaxe da questão não foi modificada, pois se considerou que a intenção da questão poderia ser diferente, uma vez que não há restrição se o aluno leu revista em quadrinhos no ano em que o questionário foi aplicado.

Comparando a Figura 6.1 com a Figura 6.4 é possível notar que a quantidade de questões isoladas ou com apenas duas utilizações é menor após a padronização. Já o número de questões presentes em todos os questionários de português que estão sendo visualizadas aumentou.

A Tabela 6.2 apresenta os resultados obtidos após a análise e padronização sintática de todos os questionários que compõem a Tabela 6.1.

Tabela 6.2. Resultados obtidos após análise e padronização sintática do questionário do SAEB.

Questionários do SAEB			Antes da análise e padronização		Depois da análise e padronização		
Nome do questionário	Nº de questionários	Nº de questões	Questões presentes em todos os questionários	Questões presentes em apenas um dos questionários	Questões com sintaxe alterada	Questões presentes em todos os questionários	Questões presentes em apenas um dos questionários
Biologia	2	177	20	137	131	37	103
Ciências	4	319	19	61	216	34	43
Diretor	5	532	3	457	784	14	286
Docentes	5	571	4	482	255	14	361
Escola	5	398	3	319	219	10	223
Física	2	177	20	137	130	36	105
Geografia	3	283	77	6	196	81	2
História	3	283	77	6	196	81	2
Matemática	15	1326	4	82	381	26	20
Português	15	1330	4	94	416	24	25
Química	2	177	20	137	128	37	103
Turma	5	138	3	43	106	18	20

Durante a realização da análise e padronização, observou-se que alguns termos usados na formulação das questões assumem significados diferentes. Por exemplo: no questionário PORTUGUES_04SERIE de 2003 existe a questão “*você tem muitos amigos na sua sala de aula?*”, em que “*sala de aula*” significa o conjunto de alunos que cursam a mesma disciplina. Ainda no questionário PORTUGUES_04SERIE de 2003, na questão “*os alunos fazem barulho e desordem*

na sua sala de aula?”, “sala de aula” significa o local físico onde os alunos assistem aulas. O mesmo significado usado na primeira questão para “sala de aula” é usado para “turma” na questão “você tem muitos amigos na sua turma?”, dos questionários PORTUGUES_08SERIE e PORTUGUES_03ANO, também de 2003.

Ao detectar essa falta de padronização na utilização dos termos, constatou-se o quanto uma ontologia seria útil para esse domínio, a fim de padronizar a utilização dos termos na formulação das questões, além de ser um recurso que permitiria inserir características adicionais à plataforma Web-PIDE, descrita no Capítulo 2.

Todos os problemas observados na base de dados do SAEB incentivaram a criação de relatórios, com o objetivo de apontar questões confusas e termos usados com significados diferentes, que são apresentados no Apêndice B.

Com todos os arquivos padronizados, foi feita a geração da LIDE e da lista de termos, lista essa que permite executar a Etapa 2 do processo ONTOP e dar início à criação de uma ontologia para o domínio de avaliações do INEP.

Inicialmente, o projeto Web-PIDE propôs a criação de uma LIDE genérica, formada pelas questões presentes em todos os questionários, e a criação das LIDES específicas de cada avaliação (SAEB, Censo Escolar, Enem e Enade). Entretanto, durante as análises do Passo 2 do processo ONTOP, notou-se que a quantidade de questões que se repetem em todos os questionários é mínima. Ao fazer uma comparação superficial com os dados do Censo Escolar (avaliação que está na fase de análise e padronização), também foi constatado que poucas questões se repetem nas avaliações, pois o enfoque de cada avaliação é diferente. Em decorrência do grande volume de dados e das dificuldades encontradas para manipular, adequar os dados para a ONTOP-Tool e padronizá-los sintaticamente, nesta pesquisa, apenas a LIDE_{SAEB} foi instanciada.

Para a geração da LIDE_{SAEB}, todos os arquivos resultantes do passo anterior estavam no mesmo diretório. O nome da LIDE foi “SAEB” e o arquivo foi salvo com o nome de “LIDE_SAEB”, no formato XML, com pode ser observado na Figura 6.5.

Na Figura 6.6 é possível observar a organização da LIDE. Por ser um arquivo extenso, a LIDE não é apresentada nesta dissertação, no entanto, pode ser encontrada em: http://www2.dc.ufscar.br/~lapes/LIDE_saeb.xml

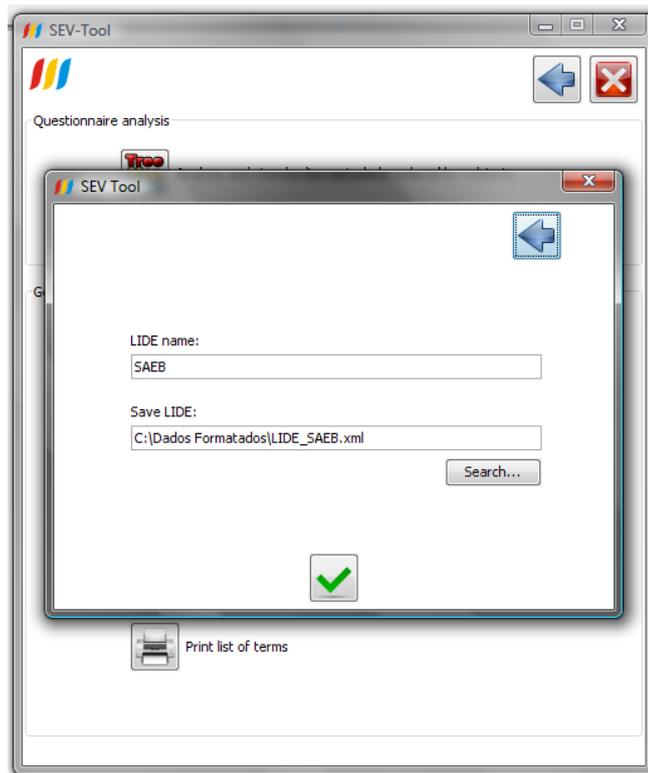


Figura 6.5. Tela para a geração da LIDE.

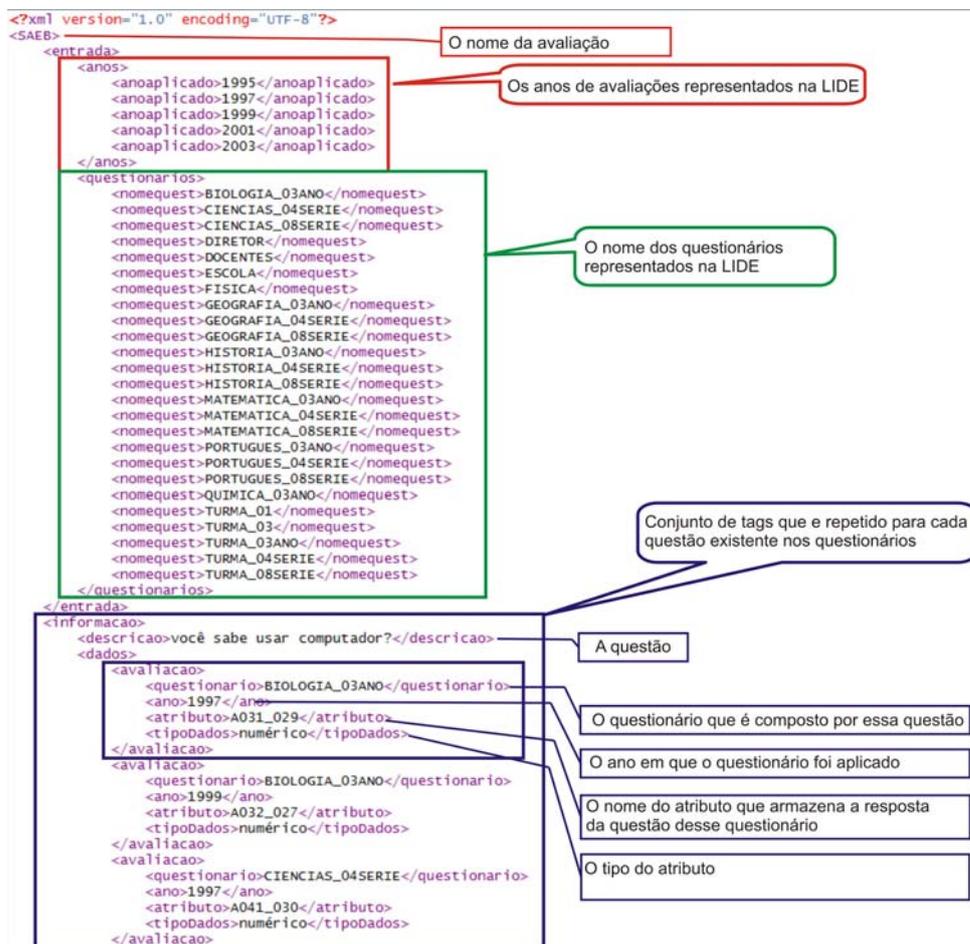


Figura 6.6. Organização da LIDE.

Após a geração da LIDE, foi gerada, por meio da *ONTOP-Tool*, a lista de termos usados nos questionários. A geração dessa lista foi feita utilizando as questões dos doze arquivos padronizados dos dados do SAEB (sessenta e seis questionários), sendo que termos comuns da língua portuguesa foram descartados por meio de uma lista pré-definida, chamada *stopwords*.

A lista, gerada em formato XML compatível com o ambiente Moodle, totalizou mil trezentos e trinta e quatro termos (1334). Esse arquivo foi importado no ambiente, como pode ser observado na Figura 6.7.



Figura 6.7. Tela do Moodle com o glossário gerado com apoio da *ONTOP-Tool*.

Com o glossário instanciado no Moodle, o próximo passo seria a inserção da definição dos termos, de acordo com o que é utilizado pelo INEP para elaborar as questões, além da exclusão de possíveis termos desnecessários ao domínio. Entretanto, estabelecer esse contato e conseguir pessoas que possam dedicar algum tempo para essa atividade é uma tarefa árdua e lenta. Por esse motivo, a definição da ontologia para avaliações educacionais do INEP, desenvolvida com o apoio do processo ONTOP, encontra-se na fase de definição do glossário.

6.3 Processo ONTOP: um exemplo com o glossário de Engenharia de Software Experimental

O glossário utilizado para exemplificar a execução do processo ONTOP foi definido pelo comitê de programa do ESELAW 2006 - *Experimental Software Engineering Latin American Workshop* (ESELAW, 2006). Esse glossário foi criado no Moodle para facilitar a comunicação entre os membros do comitê, que estavam geograficamente distribuídos. Antes de utilizar o glossário, foi feita uma atualização para verificar a necessidade de inserir novos termos sobre o domínio. Atualmente o glossário é composto por 240 termos.

A Engenharia de Software Experimental – ESE – é uma área em evidência na engenharia de software e que, define diferentes tipos de estudos experimentais, como por exemplo, *surveys*, estudos de caso e experimentos controlados, além dos procedimentos e artefatos utilizados nesse diferentes tipos de estudo (Wohlin *et al.*, 2000).

Como o objetivo desta seção é explicar o funcionamento da *Ontop-Tool* e como ela fornece apoio para a fase de conceitualização de uma ontologia, o domínio ESE não será apresentado em detalhes.

6.3.1 Passo 1: refinar o glossário de ESE

Considerando que o objetivo de usar o glossário de ESE nesta pesquisa foi decorrente da impossibilidade de usar o glossário que está sendo construído para o contexto do INEP, decidiu-se restringir o exemplo para o âmbito de Experimentos Controlados, que é um tipo de estudo experimental. Essa restrição ajuda a ilustrar o processo iterativo que pode ser realizado alternando-se entre o Moodle e a *ONTOP-Tool*. Para isso, foi necessário refinar o glossário e torná-lo mais restrito a esse contexto, facilitando a conceitualização da ontologia pretendida.

Para restringir o glossário, adotou-se o seguinte roteiro:

- 1) Exportar o glossário do Moodle em formato XML, por meio da funcionalidade disponível no próprio ambiente;
- 2) Importar o arquivo XML na *ONTOP-Tool* e visualizá-lo na ferramenta,

como pode ser observado na Figura 6.8;

- 3) Excluir termos presentes na visualização da ONTOP-Tool que não faziam parte do contexto de Experimentos Controlados;
- 4) Exportar o glossário que foi visualizado na ONTOP-Tool em formato XML compatível com o Moodle;
- 5) Importar o arquivo gerado por meio da ONTOP-Tool no ambiente Moodle, substituindo o glossário anterior;
- 6) Compartilhar o glossário com especialistas do domínio para que esses validem a exclusão e excluam novos termos, caso necessário;
- 7) Voltar à execução do item 1, até que ocorra um consenso de que o conjunto de termos represente o domínio desejado.

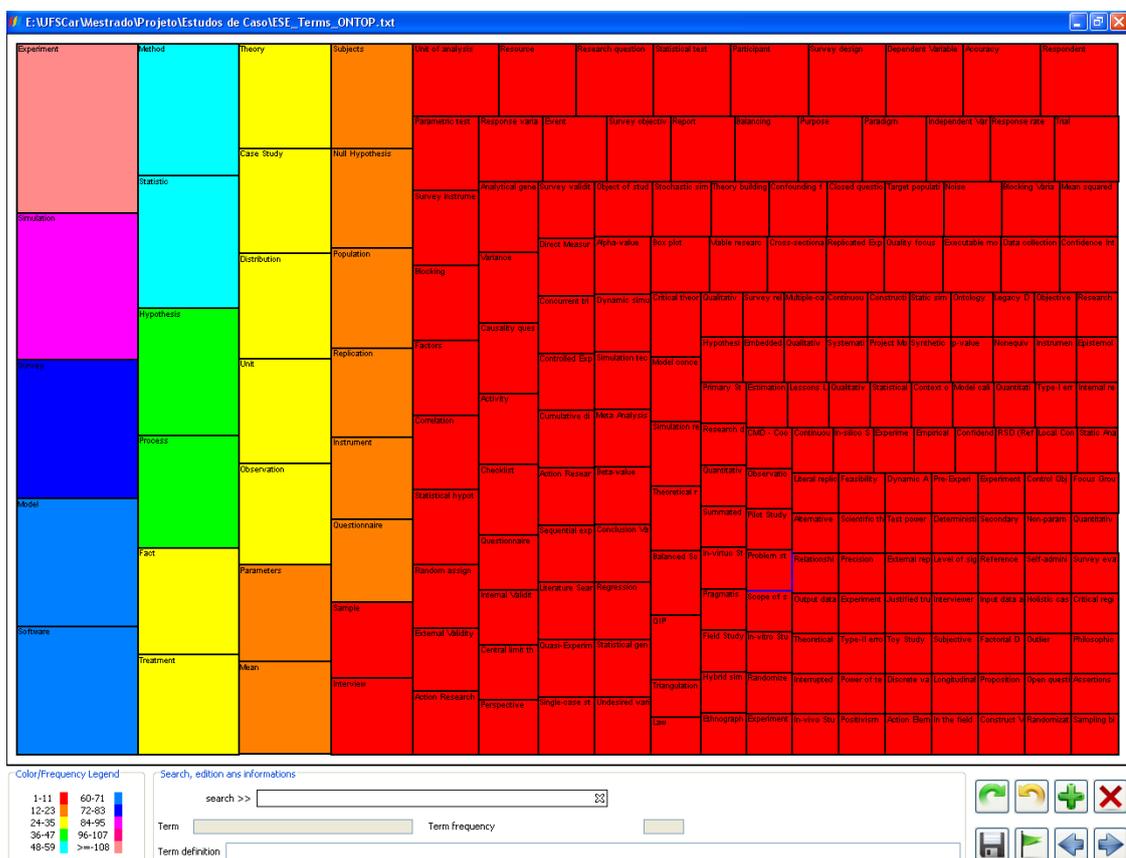


Figura 6.8. Tela da visualização inicial do glossário de ESE na ONTOP-Tool.

Na Figura 6.8 é possível observar as características da ONTOP-Tool descritas no Capítulo 5:

- a) cada caixa da representação visual representa um termo;
- b) cada caixa tem seu tamanho e sua cor definida de acordo com a

- frequência do termo, que pode ser observada na legenda, presente no canto inferior esquerdo da tela;
- c) a caixa que representa um termo já definido ou inserido na visualização atual possui a tonalidade mais clara;
- d) ao clicar na caixa é possível inserir ou editar sua definição;
- e) termos podem ser inseridos ou excluídos;

Na Figura 6.9 é possível observar a visualização do glossário com os dezoito termos obtidos após a execução do roteiro descrito acima.

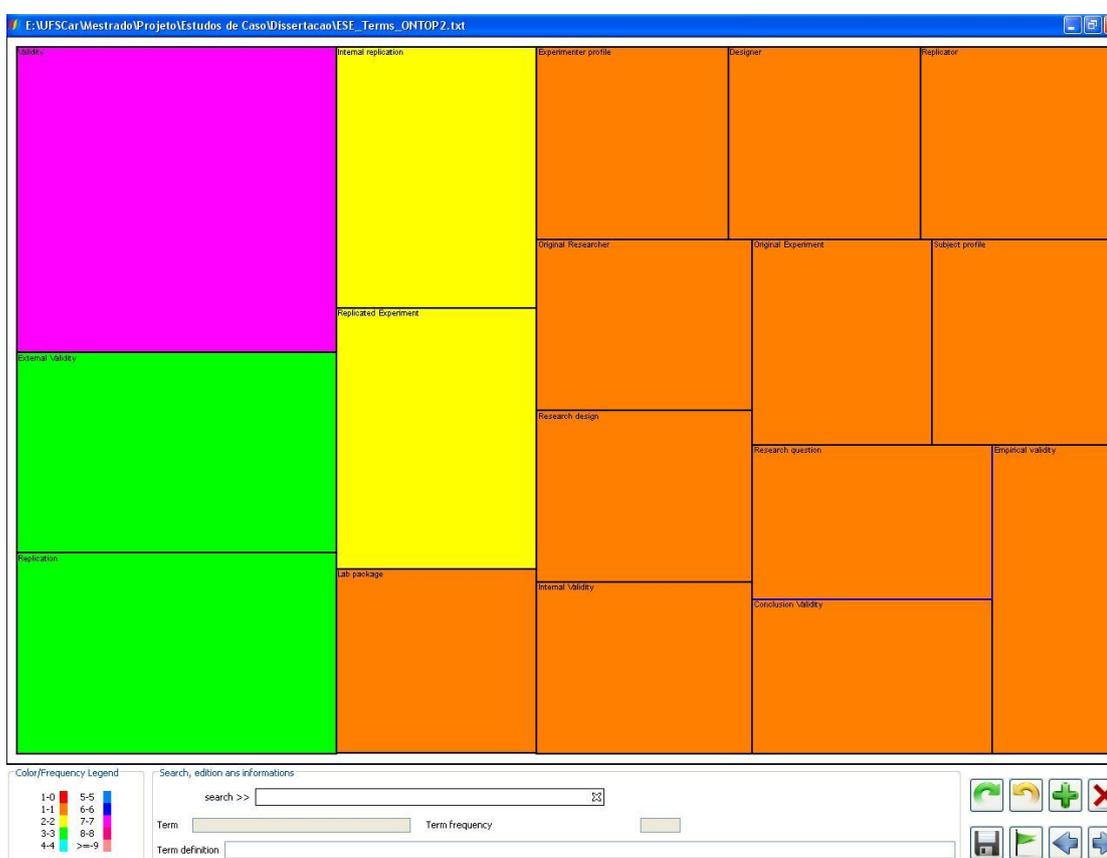


Figura 6.9. Tela da visualização do glossário de ESE na ONTOP-Tool após a adequação ao contexto de Experimentos Controlados.

6.3.2 Passo 2: definir componentes da ontologia de Experimentos Controlados

Com o glossário refinado e contendo apenas os termos pertinentes ao domínio desejado (Experimentos Controlados), foi feita a execução do Passo 2 do

processo ONTOP.

Nesse passo, os termos do glossário, que são candidatos a componentes da ontologia, foram classificados como classe, instância de uma classe (indivíduo), propriedade ou sinônimo de outro termo. Na Figura 6.10, é possível observar a tela da ONTOP-Tool para executar essa tarefa e notar que o tamanho das caixas da representação visual e suas cores destacam os termos com maior frequência de uso no glossário, os quais são fortes candidatos a se tornarem classes da ontologia.



Figura 6.10. Tela da ONTOP-Tool para definir componentes da ontologia.

Essa informação visual também auxilia o usuário a identificar, inclusive, subclasses de outra classe, fazendo buscas por palavra-chave. No exemplo em questão, ilustrado na Figura 6.10, o termo “Validity” foi selecionado (destacado na figura por um quadrado pontilhado) e definido como classe, como mostra o destaque da figura. Nesse exemplo, todos os termos foram definidos como classe.

6.3.3 Passo 3: definir hierarquia das classes

Com os componentes definidos, foi feita a execução do Passo 3 do processo

ONTOP para definir a hierarquia das classes. A ONTOP-Tool ajuda a organizá-las hierarquicamente, por meio de uma interface *drag-and-drop*.

A Figura 6.11(a) mostra a tela da ferramenta que exibe a lista dos termos do glossário que foram definidos como classes.

Na Figura 6.11(b), após ações *drag-and-drop* na ONTOP-Tool, o termo “Validity” passou a ser composto por quatro sub-classes: “Internal Validity”, “External Validity”, “Conclusion Validity” e “External Validity”, como pode ser observado no detalhe da Figura 6.11(b).

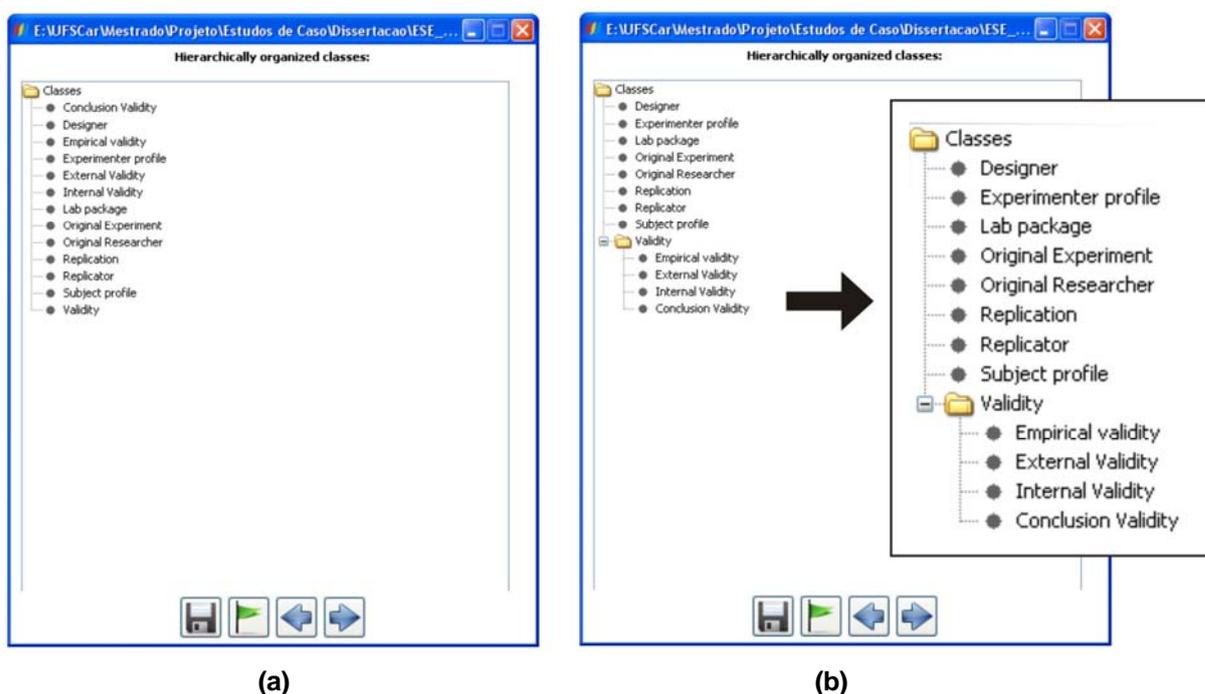


Figura 6.11. Tela da ONTOP-Tool para definição de hierarquia.

Como mencionado no Capítulo 5, esse passo do processo ONTOP não é obrigatório, pois o usuário pode usar o processo apenas para abstrair os componentes da ontologia pretendida e criar a hierarquia em uma ferramenta específica para criação e edição de ontologias.

6.3.4 Passo 4: definir relacionamento entre as classes

Definidas as classes no Passo 2, a ONTOP-Tool ajuda a definir os domínios (*Domains*) e escopos (*Ranges*) das propriedades, detalhados no Capítulo 4.

Como pode ser observado na Figura 6.12, nessa interface da ferramenta as classes definidas no Passo 2 são apresentadas do lado direito e do lado esquerdo da

tela e, entre elas, coloca-se à disposição do usuário uma lista de propriedades pré-definidas na ferramenta ou definidas no Passo 2. Também é permitido que o usuário adicione, quando necessário, novas propriedades. Para estabelecer o *Domain-Range* de uma propriedade, o usuário deve:

- a) selecionar uma classe da lista da esquerda, que será o *Domain* (domínio) ;
- b) selecionar uma propriedade na lista do centro;
- c) selecionar uma classe da lista da direita; que será o *Range* (escopo);
- d) confirmar o relacionamento.

No exemplo da ontologia para Experimentos Controlados, foram criados dezessete relacionamentos, sendo que alguns estão destacados na Figura 6.12. Na listagem dos relacionamentos criados, as propriedades estão sublinhadas:

- 1) *Validity* contains *Internal Validity*;
- 2) *Validity* contains *External Validity*;
- 3) *Validity* contains *Conclusion Validity*;
- 4) *Validity* contains *Empirical validity*;
- 5) *Subject Profile* influences *Validity*;
- 6) *Original Researcher* has *Experimenter profile*;
- 7) *Original Researcher* conducts *Original Experiment*;
- 8) *Original Experiment* describes *Lab package*;
- 9) *Lab package* is basis for *Replication*;
- 10) *Replication* has *Validity*;
- 11) *Replication* generates *Lab package*;
- 12) *Replicator* conducts *Replication*;
- 13) *Replicator* has *Experimenter profile*;
- 14) *Experimenter profile* is recorded *Lab package*;
- 15) *Subject* execute *Replication*;
- 16) *Subject* has *Subject profile*;
- 17) *Designer* designs *Original Experiment*.

Assim como o Passo 3, esse passo do processo ONTOP pode não ser executado, caso o usuário prefira definir os domínios e escopos em uma ferramenta

específica para criação e edição de ontologias.

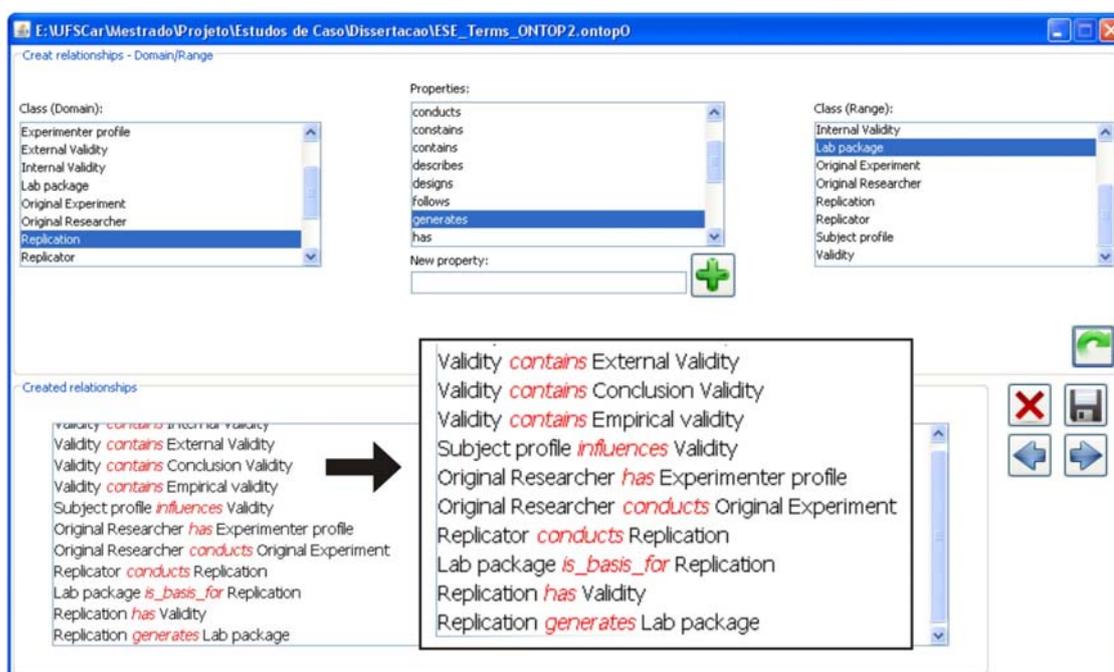


Figura 6.12. Tela da ONTOP-Tool para definir *Domain-Range* das propriedades.

6.3.5 Passo 5: gerar arquivo OWL

Finalmente, toda a definição conceitual da ontologia, norteadas pelo processo ONTOP e facilitada com o apoio computacional da ONTOP-Tool, pode ser transferida para um arquivo OWL e então importada em diversas ferramentas que dão suporte à criação e edição de ontologias.

A ferramenta utilizada nesta pesquisa é a Protégé-2000 versão 3.4, na qual o arquivo OWL, relativo à ontologia de Experimentos Controlados que foi iniciada com o processo ONTOP, foi importado. Na Figura 6.13 é possível observar as classes e, na Figura 6.14, as propriedades, definidas com apoio da ONTOP-Tool, agora instanciadas na ferramenta.

Na Figura 6.15 é apresentada a visualização do arquivo OWL por meio do *plug-in* Jambalaya (2009), na qual é possível observar, por meio das setas criadas na visualização, o relacionamento entre a classe definida como domínio e a classe definida como escopo de uma determinada propriedade, lembrando que esse relacionamento foi definido no Passo 4 do processo, com o apoio da ONTOP-Tool.

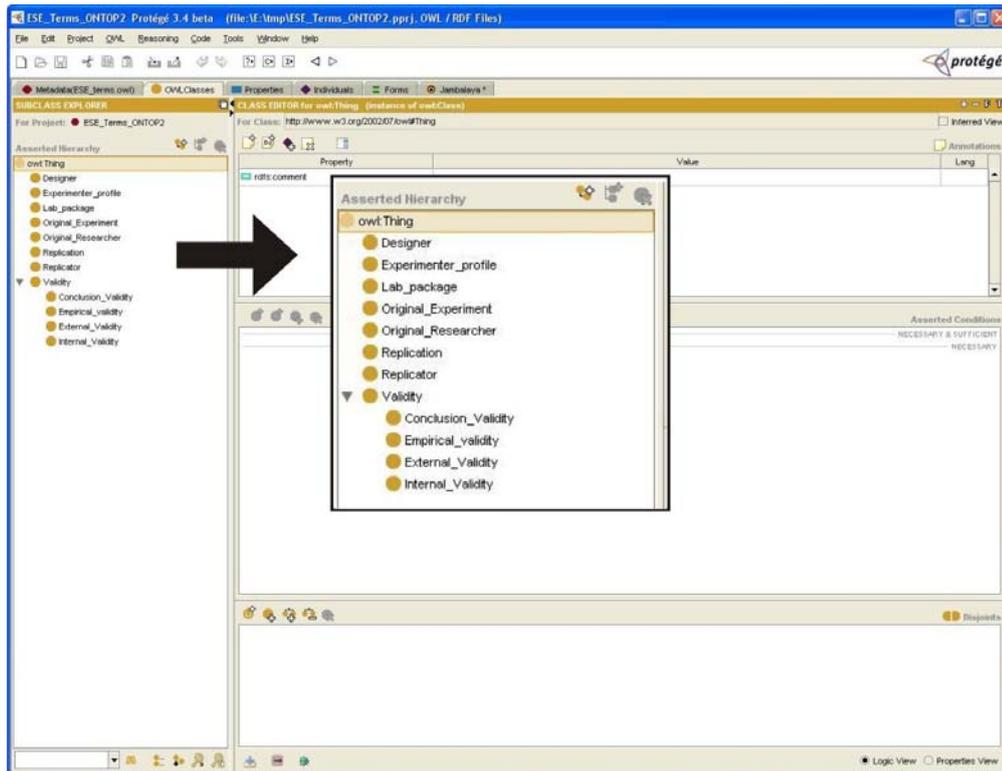


Figura 6.13. Tela da Protégé-2000 com as classes instanciadas.

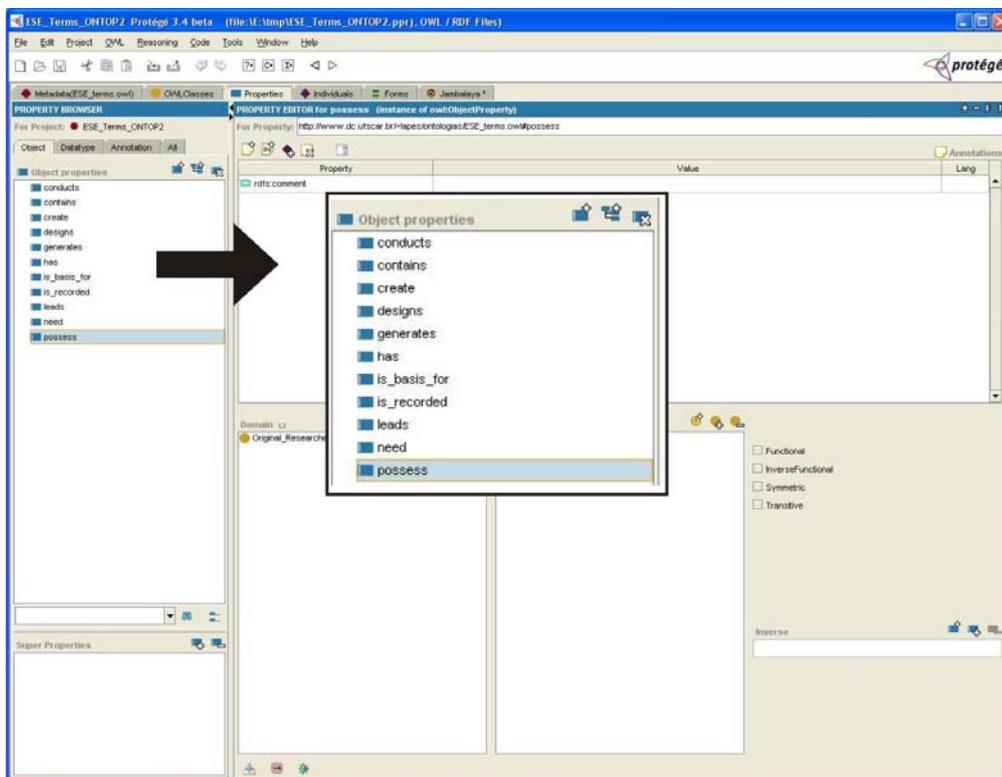


Figura 6.14. Tela da Protégé-2000 com as propriedades instanciadas.

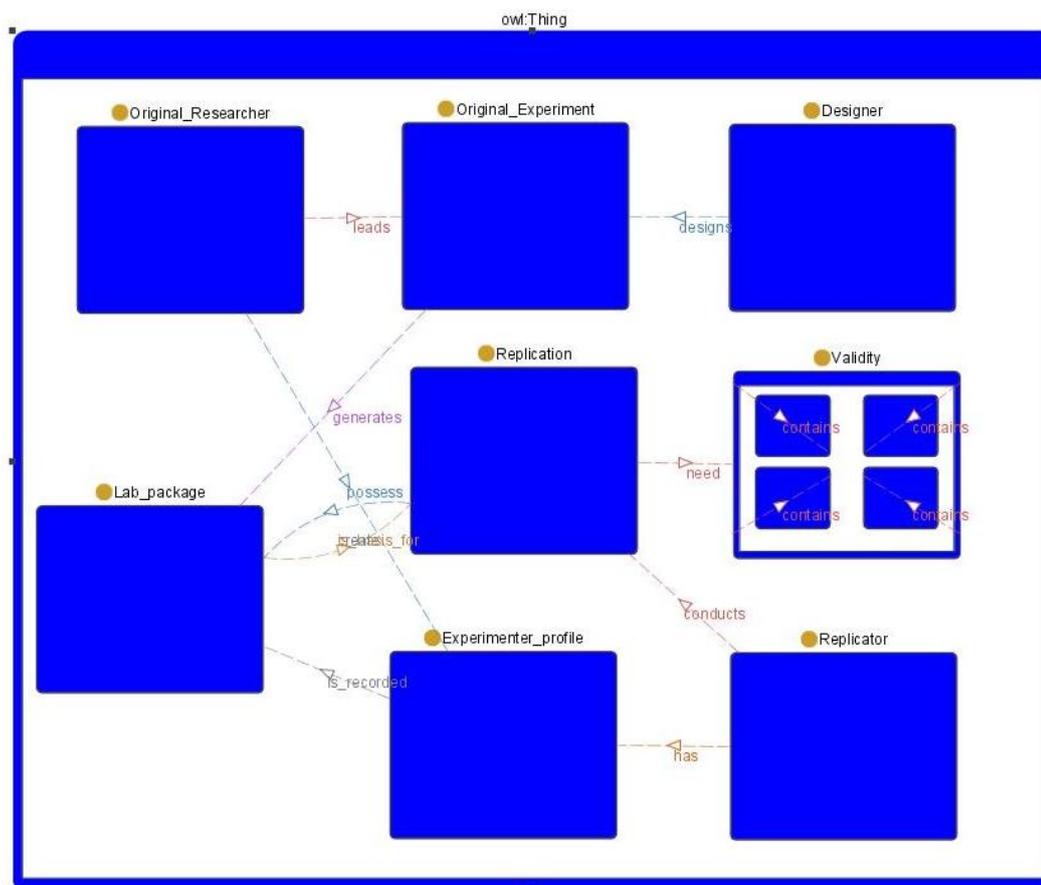


Figura 6.15. Visualização do arquivo OWL por meio do plug-in Jambalaya (2009).

Com o arquivo OWL importado em uma ferramenta adequada, os próximos passos para definição de uma ontologia podem ser executados com base nessa conceitualização inicial do domínio realizada com o apoio do processo ONTOP e da *ONTOP-Tool*.

Certamente é indispensável a participação de especialistas do domínio para que o resultado final do processo ONTOP represente de maneira fiel o domínio para o qual a ontologia está sendo criada.

6.4 Considerações Finais

Neste capítulo exemplificou-se a aplicação dos processos P-LIDE e ONTOP com o apoio das ferramentas *SEV-Tool* e *ONTOP-Tool*, respectivamente.

A aplicação do P-LIDE resultou na LIDE_{SAEB} e em um glossário com termos usados para compor os questionários dessa avaliação. Esse glossário está em

processo de refinamento para que os passos para iniciar a criação de uma ontologia para esse domínio possam ser executados.

Os problemas inicialmente detectados nos dados do INEP fizeram com que os esforços para a construção da LIDE se concentrassem, inicialmente, na avaliação SAEB. No entanto, o processo P-LIDE e a *SEV-Tool* permitem que os dados de todas as avaliações do INEP sejam analisados, padronizados e tenham a respectiva LIDE gerada. A contribuição desta pesquisa, nesse contexto, foi identificar uma abordagem que permitisse analisar os dados das avaliações educacionais do INEP. Assim, para os dados das outras avaliações, deve-se repetir o processo aqui exemplificado com o SAEB.

A aplicação do processo ONTOP com o apoio da *ONTOP-Tool*, nesta pesquisa, resultou em uma pré-conceitualização de uma ontologia sobre Experimentos Controlados. O uso da visualização para evidenciar ao usuário termos mais usados nas definições, auxiliou a abstração dos componentes da ontologia. Com o uso das outras funcionalidades da ferramenta, novas informações foram abstraídas e, ao final do processo, se obteve um arquivo OWL.

Ao gerar um arquivo OWL automaticamente pela *ONTOP-Tool* e importá-lo na Protégé-2000, as próximas fases da construção da ontologia (criação de axiomas, por exemplo) podem ser imediatamente iniciadas, poupando trabalho de criação e instanciação de todas as classes e relacionamentos.

Capítulo 7

CONCLUSÕES

Este capítulo apresenta as conclusões da pesquisa realizada nesta dissertação, evidenciando as contribuições, as limitações e possíveis trabalhos futuros.

Nesta dissertação foram propostos os processos P-LIDE e ONTOP e as ferramentas *SEV-Tool* e *ONTOP-Tool* que apóiam a execução desses processos. A pesquisa que resultou nesses produtos foi motivada pelo projeto Web-PIDE do Observatório da Educação.

Com o intuito de criar uma plataforma web para compartilhar e integrar os dados educacionais oriundos de dezoito anos de avaliações do INEP, a plataforma visa facilitar a consulta e geração de hipóteses relativas a esses dados e minimizar o problema com as bases de dados do INEP. Dessa forma, tais informações podem se tornar acessíveis a gestores educacionais, diretores de escola, professores, pais de alunos e também serem integradas com outros sistemas de informação.

Na arquitetura dessa plataforma, foi definido que a base para o seu desenvolvimento seria uma linguagem de marcação intitulada LIDE, que foi a motivação inicial desta pesquisa. A LIDE contém informações das questões presentes em cada avaliação educacional do INEP, indicando os anos em que as questões foram utilizadas e o atributo na base de dados do INEP que representa cada questão. Durante o desenvolvimento da LIDE, problemas foram identificados, o que motivou a investigação de soluções para os mesmos, resultando no processo P-LIDE e no apoio computacional *SEV-Tool* propostos neste trabalho.

Embora houvesse um conjunto grande de avaliações, no contexto deste trabalho a manipulação dos dados se limitou ao SAEB, por solicitação do próprio INEP.

Ao iniciar o planejamento da LIDE, notou-se a falta de padronização da formatação dos dados e de incompatibilidade entre os arquivos. Após se decidir pelo formato de arquivo SAS, verificou-se que seria necessária uma padronização sintática das avaliações, pois muitas questões são iguais semanticamente, mas têm suas sintaxes diferentes, o que comprometia a criação da LIDE.

O P_LIDE prevê a análise e a padronização sintática das questões que compõem os questionários do INEP, provendo na *SEV-Tool* uma funcionalidade específica para essa tarefa que, por estar apoiada em visualização, torna a tarefa menos árdua. Feita essa padronização é possível gerar a LIDE correspondente à avaliação que está sendo tratada no momento.

Ao tentar gerar a LIDE para os dados das avaliações educacionais do SAEB, identificou-se que os termos usados nas questões eram empregados com significados distintos em algumas questões.

Atentando-se para a necessidade de se ter um vocabulário compartilhado para esse domínio juntamente com o objetivo do projeto Web-PIDE de compartilhar um grande volume de dados por meio de uma plataforma web, verificou-se que a criação de uma ontologia poderia minimizar o problema de inconsistência dos termos. Além disso, ontologia é um recurso comumente utilizado no contexto de Web Semântica, que tem como alvo de pesquisa mecanismos apropriados para o compartilhamento de grande volume de dados na web.

Ao decidir-se pelo desenvolvimento de uma ontologia, pesquisaram-se métodos que dessem suporte a essa tarefa e detectou-se a falta de apoio computacional para a fase de Conceitualização, que está presente na grande maioria dos métodos encontrados na literatura, além do método Methontology, considerado nesta pesquisa.

Assim, no processo ONTOP, são previstas as atividades de refinamento de um glossário, abstração dos componentes da ontologia, hierarquização das classes, estabelecimento de alguns relacionamentos e por fim, a geração do arquivo OWL. As duas primeiras atividades têm suas respectivas funcionalidades da *ONTOP-Tool* auxiliadas pela visualização, que considera a frequência de cada termo para dispor os dados na tela para o usuário.

Com um arquivo OWL representando todas as informações adquiridas com apoio do processo ONTOP e da *ONTOP-Tool*, o usuário pode importá-lo em uma ferramenta apropriada para a execução das outras fases do desenvolvimento de uma ontologia e, imediatamente, iniciar as definições de outros componentes.

Em resumo, este trabalho apresentou dois processos que, apesar de terem objetivos específicos, se relacionam entre si. No P-LIDE é possível realizar o tratamento nos dados das avaliações educacionais do INEP com a finalidade de gerar a LIDE. Ao final da aplicação desse processo também é possível gerar automaticamente, por meio da *SEV-Tool*, um glossário com termos utilizados para compor as questões que estão representadas na LIDE. No ONTOP geram-se informações iniciais para a definição de uma ontologia, as quais dão suporte à fase de conceitualização de engenharia de ontologia. A entrada para esse processo é um glossário que pode ser derivado do P-LIDE, caso se esteja trabalhando com os dados do INEP, ou pode ser criado no Moodle (Moodle, 2009), especificamente para o processo ONTOP, sem estar vinculado ao contexto do INEP. Ambos os processos são apoiados por recursos de visualização, o que facilita a análise humana, indispensável em todos os passos.

Assim, pode-se considerar que um dos pontos principais nesta pesquisa é o uso da visualização tanto para padronização de metadados (dados do INEP), quanto para auxiliar no desenvolvimento de ontologias, considerando, nessa última atividade, o refinamento de glossários e a abstração de componentes.

7.1 Contribuições e limitações

Esta pesquisa traz contribuições tanto para o âmbito das avaliações educacionais do INEP, como para o desenvolvimento de ontologias.

Em relação às contribuições para as avaliações educacionais do INEP, pode-se pontuar:

- A estratégia baseada em visualização para tratamento dos dados relativos às avaliações do INEP;

- O apoio computacional provido pela ferramenta para a análise e comparação dos questionários das avaliações, utilizando visualização;
- O processo sistematizado para geração das LIDEs;
- A padronização sintática dos dados do SAEB e a LIDE|_{SAEB};
- O relatório com indicação de possíveis modificações nas questões para que se obtenha uma maior padronização e facilidade nas análises, sendo que as indicações podem ser válidas para outras avaliações;
- A geração automática de uma lista de termos usados para compor as questões, sendo que essa lista se torna um glossário ao ser importada no ambiente Moodle, o que permite dar continuidade ao processo de desenvolvimento de uma ontologia para o domínio de avaliações do INEP.

Em relação à contribuição mencionada acima, sobre o apoio computacional provido pela *SEV-Tool* para a análise e comparação dos questionários, ressalta-se que essa característica pode ser um auxílio para a descoberta de tendências, similaridades e diferenças, caso ocorra uma comparação entre os dados do Enem armazenados pelo INEP desde 1998 com a nova estrutura da avaliação proposta pelo governo para ser implantada em 2009.

Essa mesma contribuição pode ter uma amplitude maior se disponibilizada para pesquisadores de outras áreas, principalmente os que não têm o enfoque na área computacional, mas que necessitam analisar e abstrair tendências e informações implícitas sobre os dados das avaliações educacionais do INEP. Vale ressaltar que em 2006 foram aprovados vinte e oito projetos e em 2008, vinte e nove.

Em relação às contribuições para o desenvolvimento e ontologias, podem-se pontuar:

- O apoio computacional fornecido pela *ONTOP-Tool* e a estratégia definida por meio do processo ONTOP para a fase de Conceitualização do método Methontology;
- O uso de glossários para iniciar o desenvolvimento de ontologias, o que pode ser considerado também como uma forma de documentação;

- O uso da visualização (na *ONTOP-Tool*) para auxiliar o refinamento de glossários;
- O uso do ambiente web Moodle para gerir os glossários, o que possibilita o compartilhamento com outras pessoas que possam agregar valor ao desenvolvimento da ontologia;
- A investigação do uso da visualização para auxiliar a abstração dos componentes da ontologia;
- A geração do arquivo OWL, que permite carregar as informações definidas durante a execução do processo ONTOP em uma ferramenta específica para criação e edição de ontologias;

Apesar das contribuições mencionadas, a pesquisa apresenta algumas limitações, como por exemplo:

- Implementação somente da técnica de visualização *Tree-Map*;
- A *SEV-Tool* e a *ONTOP-Tool*, por consequência do uso da técnica *Tree-Map*, comprometem a visualização de um grande conjunto de questões ou termos, no caso do glossário, fazendo com que alguns retângulos tenham um tamanho muito reduzido, difícil de serem distinguidos. O número ideal de dados para a visualização depende da resolução da tela do computador usado, da quantidade de questões (termos) agrupadas e da quantidade de questionários em que as questões estão divididas;
- Ao final do processo, é permitido gerar apenas arquivos OWL, não havendo opções para arquivos RDFs, por exemplo;
- A *SEV-Tool* está preparada para importar apenas metadados no formato definido no Capítulo 6, que condiz com os metadados disponibilizados pelo INEP. Para visualizar metadados de outro contexto, ou é necessário adequá-los ao formato definido, ou customizar a ferramenta.

7.2 Trabalhos futuros

Entre as atividades que podem dar continuidade a esta pesquisa, podem-se citar:

- Definir a LIDE|_{CENSOESCOLAR}, LIDE|_{ENADE}, e LIDE|_{ENEM};
- Finalizar a ontologia para Experimentos Controlados, com criação de axiomas, indivíduos, etc;
- Finalizar o glossário sobre avaliações educacionais do INEP;
- Definir uma ontologia para avaliações educacionais do INEP;
- Inserir na *SEV-Tool* e na *ONTOP-Tool* processamentos linguísticos para classificação sintática, semântica e morfológica, a fim de aprimorar a extração de termos na geração da lista de termos (processo P-LIDE) e, principalmente, para auxiliar o usuário a abstrair os componentes da ontologia no Passo 2 do processo ONTOP;
- Assim que a plataforma Web-PIDE for desenvolvida, validar a LIDE com situações reais para mostrar sua utilidade e auxílio a buscas nas bases de dados do INEP;
- Adicionar um módulo na *SEV-Tool* para gerir a lista *stopwords*;
- Adicionar um módulo na *SEV-Tool* para formatar metadados, inserindo as tabulações adequadas, assim como um módulo que possibilitasse adaptar a visualização para diferentes formatos de metadados;
- Complementar o processo ONTOP e a *ONTOP-Tool* para darem apoio a outras atividades no desenvolvimento de ontologia, como por exemplo, a Garantia de Qualidade, Documentação e Validação;
- Integrar à *SEV-Tool* e à *ONTOP-Tool* novas técnicas de visualização para que o usuário tenha outras possibilidades;
- Inserir na *ONTOP-Tool* uma funcionalidade para exportar a conceitualização realizada em outros formatos, como o XMI (*XML Metadata Interchange*), para que seja possível manipular as informações em ferramentas de modelagem UML.

REFERÊNCIAS

ALMEIDA, Mauricio B.; BAX, Marcello P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, Brasília, v. 32, n. 3, p. 7-20, dez., 2003.

ARPÍREZ, Julio C; CORCHO, Oscar; FERNÁNDEZ-LÓPEZ, Mariano; GÓMEZ-PÉREZ, Asunción. WebODE: a scalable workbench for ontological engineering. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE, K-CAP, 2001, Victoria, British Columbia. **Proceedings...** New York: ACM Publisher, 2001. p. 6-13.

AUVIL, Loretta; LLORÀ, Xavier; SEARSMITH, Duane; SEARSMITH, Kelly. VAST to Knowledge: Combining tools for exploration and mining. **IEEE Symposium on Visual Analytics Science and Technology - VAST**. Philadelphia, p. 197-198, Nov., 2007.

BASÉGIO, Túlio. **Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na Língua Portuguesa do Brasil**. Dissertação (Mestrado em Computação Aplicada) - Pontifícia Universidade Católica do Rio Grande do Sul, São Leopoldo, 2006.

BECKER, Barry. G. Using MineSet for knowledge discovery. **IEEE Computer Graphics and Applications**. Mountain View, v. 17, n. 4, p. 75-78, Jul./Aug., 1997.

BEDERSON, Benjamin B. Fisheye menus. In: ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY - UIST, 13th, 2000, San Diego, . **Proceedings...** New York: ACM Publisher, 2000. p. 217-225.

BIOML. **Biological Markup Language**. Disponível em: <<http://xml.coverpages.org/bioml.html>>. Acesso em: Jun. 2009.

BLACKBURN, Simon. **Dicionário Oxford de Filosofia**. São Paulo, Brasil: Jorge Zahar, 1997. 437p.

CEMIN, Cristina. **Visualização de informações aplicada à gerência de software**. 70f. Dissertação (Mestrado Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

CHEN, Chaomei; KULJIS, Jasna; PAUL, Ray J. Visualizing latent domain knowledge. **IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews**. Philadelphia, v. 31, n. 4, p. 518-529. Nov. 2001.

CHERNOFF, Herman. The Use of Faces to Represent Points in K-Dimensional Space Graphically. **Journal of the American Statistical Association**, Alexandria, v. 68, n. 342, p. 361-368, Jun. 1973.

CONNOLLY, Dan; HARMELEN, Frank van; HORROCKS, Ian; MCGUINNESS, Deborah L.; PATEL-SCHNEIDER, Peter F.; STEIN, Lynn Andrea. DAML+OIL Reference Description. Cambridge: W3C, Março, 2001. Disponível em: <<http://www.w3.org/TR/daml+oil-reference>>. Acesso em: Jun. 2009. Relatório Técnico.

DACONTA, Michael. C., OBRST, Leo. J., SMITH, Kevin. T.: **The Semantic Web: A guide to future of XML, Web Services and Knowledge Management**. Indianapolis: Wiley Publishing, 2003. 281p.

DATAESCOLA. **Portal de consulta dos dados da Educação Básica**. Disponível em: <<http://www.inep.gov.br/censo/basica/dataescolabrasil>>. Acesso em: mai. 2009.

DEVEDŽIĆ Vladan. Understanding Ontological Engineering. **Communications of the ACM archive: Supporting community and building social capital - Special Issue: Virtual extension table of contents**. New York, v. 45, n. 4, p. 136-144, Apr. 2002.

DOMINGUE, John. Tadzebao and WebOnto: discussing, browsing and editing ontologies on the web. In: BANFF KNOWLEDGE ACQUISITION, MODELING AND MANAGEMENT WORKSHOP. 11th, 1998, Alberta. **Proceedings...** Banff, KAW Press, 1998.

EDUDATABRASIL. **Portal do Sistema de Estatísticas Educacionais**. Disponível em: <<http://www.edudatabrasil.inep.gov.br>>. Acesso em: mai. 2009.

ESEGLOSSARY. **Glossário de Engenharia de Software Experimental**. Disponível em: <http://lens-ese.cos.ufrj.br/wikiese/index.php/Simulation_technique>. Acesso em: Jun. 2009.

ESELAW. Glossário de Engenharia de Software Experimental. In: EXPERIMENTAL SOFTWARE ENGINEERING LATIN AMERICAN WORKSHOP, 2006, Rio de Janeiro. **Proceedings...**, Rio de Janeiro: SBC, 2006, pp. 1-33.

FALBO, Ricardo de Almeida, MENEZES, Crediné Silva de; ROCHA, Ana Regina C. A systematic approach for building ontologies. In: IBERO-AMERICAN CONFERENCE ON ARTIFICIAL INTELLIGENCE - IBERAMIA. 6th, 1998, London, UK. **Proceedings...** London: Springer-Verlag, 1998, p. 346-360.

FENSEL, Dieter. **Ontologies: silver bullet for knowledge management and electronic commerce**. London: Springer-Verlag, 2001. 138p.

FENSEL Dieter; HARMELEN, Frank van. Project Presentation On-To-Knowledge: Content-driven Knowledge-Management Tools through Evolving Ontologies. Amsterdam: Vrije Universiteit Amsterdam, Jan. 1999.

Disponível em: <<http://www.ontoknowledge.org/down/del33.pdf>>. Acesso em: Jun. 2009. Relatório Técnico.

FENYÖ, David. The Biopolymer Markup Language. **Bioinformatics**, Oxford, v.15, n. 4, p. 339-340, Apr. 1999.

FERNÁNDEZ, M.; GOMÉZ-PÉREZ, Asunción; JURISTO, Natalia. Methontology: from ontological art towards ontological engineering. In: SPRING SYMPOSIUM SERIES ON ONTOLOGICAL ENGINEERING – AAAI, 1997, Stanford. **Proceedings...** California: AAAI Press, p. 33-40.

FERNÁNDEZ-LÓPEZ, Mariano, GÓMEZ-PÉREZ Asunción, SIERRA, Juan Pazos; SIERRA, Alejandro Pazos. Building a Chemical Ontology Using Methontology and the Ontology Design Environment. **IEEE Intelligent Systems & their applications**, Maryland, v. 4, n. 1, p.37-46, Jan./Feb. 1999.

FISHEYEMENU. **Fish-Eye Menu demo** Disponível em: <<http://www.cs.umd.edu/hcil/fisheyemenu/fisheyemenu-demo.shtml>>. Acesso em: Jun. 2009.

FREITAS, Carla Maria Dal Sasso; WAGNER, Flávio Rech. A Methodology for Selecting Visual Representations in Scientific and Simulation Applications. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO GRÁFICA E PROCESSAMENTO DE IMAGENS, 6. SIBGRAPI, 1993, Recife. **Anais...** Recife: SBC/UFPE, 1993. p. 89-97.

FREITAS, Carla Maria Dal Sasso; CHUBACHI, Olinda Mioka; LUZZARDI, Paulo Roberto Gomes; CAVA, Ricardo Andrade. Introdução à Visualização de Informações. **RITA - Revista de Informática Teórica e Aplicada**. Porto Alegre, v. 8, n. 2, p. 143-158, 2001.

FURNAS, George W., The FISHEYE view: A new look at structured files. Madison: Bell Laboratories n. 82, Out. 1982. Relatório Técnico.

GERSHON, Nahum.; EICK, Stephen.G., CARD, Stuart. Information visualization interactions. **ACM Interactions**, Nova Iorque, v. 5, n. 2, p. 9-15, Mar./Apr. 1998.

GIL, Jose Nilton; HIRATA, Celso M. XACDML- Extensible ACD Markup Language. In: ANNUAL SYMPOSIUM ON SIMULATION - ANSS, 36., 2003, Washington. **Proceedings...** Washington, EUA: IEEE Computer Society, 2003. p. 343-350

GÓMEZ-PÉREZ Asunción, FERNÁNDEZ-LÓPEZ Mariano, CORCHO Oscar **Ontological Engineering: with examples from the areas of knowledge management, e-commerce and the Semantic Web**. 2ª Ed. London: Springer-Verlag, 2003. 362p.

GÓMEZ-PÉREZ, Assunción; MANZANO-MACHO, David. A survey of Ontology Learning Methods and techniques: OntoWeb Deliverable 1.5. Madrid: Universidad Politécnica de Madrid, Mai. 2003. Disponível em: <<http://www.sti-innsbruck.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf>>. Acesso em: Jun. 2009. Relatório Técnico.

GRAPHVIZ. **Graph Visualization Software**. Disponível em: <<http://www.graphviz.org>>. Acesso em: Jun. 2009.

GRAVES, Mark. **Projeto de Banco de Dados com XML**. São Paulo: Makron Books, 2003. 534p.

GRUBER, Thomas R. A Translation Approach to Portable Ontology Specifications. **Knowledge Acquisition**, Stanford, v. 5, n. 2, p. 199-220, Jun. 1993.

GRUBER, Thomas R. Towards Principles for the Design of Ontologies used for Knowledge Sharing, **International Journal of Human-Computer Studies**, Stanford, v. 43, n. 5-6, p. 907-928, Nov. 1995

GRUNINGER, Michael, FOX, Mark. S. Methodology for the Design and Evaluation of Ontologies. In: WORKSHOP ON BASIC ONTOLOGICAL ISSUES IN KNOWLEDGE SHARING, 1995. Montreal. **Proceedings...** Montreal: AAAI Press, 1995. p. 1-10.

HOLSAPPLE, Clyde W.; JOSHI, Kshiti D., A collaborative approach to ontology design. **ACM Communications**. New York, v. 45, n. 2, p. 42-47, Feb. 2002.

HORRIDGE, Matthew; KNUBLAUCH, Holger; RECTOR, Alan; STEVENS, Robert; WROE, Chris. A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools. Manchester; University of Manchester, Aug. 2004. Tutorial.

HORROCKS, Ian; FENSEL, Dieter; HARMELEN, Frank Van; DECKER, Stefan.; ERDMANN, Michael; KLEIN, Michel C. A. OIL in a Nutshell. In: WORKSHOP ON APPLICATION OF ONTOLOGIES AND PSMs - ECAI, 2000, Berlim. **Proceedings...** Berlin: Editora, 2000. p. 4.1-4.12.

HYPERGRAPH. **Graph visualization using hyperbolic geometry**. Disponível em: <<http://hypergraph.sourceforge.net/>>. Acesso em: Jun. 2009.

ICHISE, Ryutaro; SATOH, Kazuhiro; NUMAO, Masayuki. Elucidating Relationships among Research Subjects from Grant Application Data. **Information Visualization**, New York, p. 427-432. Jul. 2008.

INEP - Instituto **Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**. Disponível em: <<http://www.INEP.gov.br>>. Acesso em: jan. 2009.

INSELBERG, Alfred; DIMSDALE, Bernard. Parallel coordinates: a tool for visualizing multidimensional geometry. In: CONFERENCE ON VISUALIZATION - VIS, 1990, Los Alamitos. **Proceedings...** New York: IEEE Computer Society, 1990. p. 361–378.

JAMBALAYA. **Protégé plug-in to visualize the knowledge bases**. Disponível em: <<http://www.thechiselgroup.org/jambalaya>>. Acesso em: mai. 2009.

JFLAP. **Software for experimenting with formal languages topics**. Disponível em: <<http://www.jflap.org/>>. Acesso em: jun. 2009.

JGRAPHT. **Free Java graph library**. Disponível em: <<http://jgrapht.sourceforge.net/>>. Acesso em: Jun. 2009.

JOHNSON, Brian; SHNEIDERMAN, Ben. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In: CONFERENCE ON VISUALIZATION - VIS, 2., 1991, San Diego. **Proceedings...** New York: IEEE Computer Society, 1991. p. 284–291.

KARP, Peter D.; CHAUDHRI, Vinay K.; THOMERE, Jerome. XOL: An XML-Based Ontology Exchange Language. Version 0.5. California: IA Center, SRI International n. 559, Feb. 2000. Disponível em: <www.ai.sri.com/~pkarp/xol>. Acesso em: Jun. 2009. Relatório Técnico.

KARVE, Aneesh; GLEICHER, Michael. Glyph-based Overviews of Large Datasets in Structural Bioinformatics. In: INTERNATIONAL CONFERENCE ON INFORMATION VISUALIZATION, 11., 2007, Zurich, SW. **Proceedings...** New York: IEEE Computer Society, 2007. p.1-6.

KATIFORI, Akriv; HALATSIS, Constantin; LEPOURAS, George; VASSILAKIS, Costas; GIANNOPOULOU, Eugenia. Ontology visualization methods - a survey. **ACM Computing Surveys**, New York, v. 39, n. 4, p. 10-53, Oct. 2007

KISHORE, Rajiv; ZHANG, Hong; RAMESH, Ramaswamy. A Helix-Spindle model for ontological engineering. **ACM Communications**, New York, v. 47, n. 2, p. 69-75, Feb. 2004.

LAMPING, Jonh.; RAO, Ramana. The hyperbolic browser: a focus+context technique for visualizing large hierarchies. **Journal of Visual Languages and Computing**. Califórnia, v. 7, n.1, p.33-55, Mar. 1996.

LASSILA, Ora.; SWICK, Ralph R. Resource Description Framework (RDF) model and syntax specification. W3C Recommendation. 1998. Disponível em: <<http://www.w3.org/TR/REC-rdf-syntax/>>. Acesso em: Jun. 2009.

LINSTONE, Harold A.; TUROFF, Murray. The Delphi Method Techniques and Applications. London: Addison-Wesley, 1975. 621p.

LIVEPLASMA. **Informações sobre música e cinema**. Disponível em: <<http://www.musicplasma.com>>. Acesso em: jun. 2009.

LUKE, Sean; HEFLIN, Jeff. SHOE 1.01, Proposed Specification. Maryland: Department of Computer Science, University of Maryland, Apr. 2000. Disponível em: <<http://www.cs.umd.edu/projects/plus/SHOE/spec.html>>. Acesso em: Jun. 2009. Relatório Técnico.

LUZZARDI, Paulo Roberto Gomes. **Crêterios de Avaliaçãõ de Têcnicas de Visualizaçãõ de Informações Hierárquicas**. Tese (Doutorado em Ciênciã da Computaçãõ) - Instituto de Informãtica, Universidade Federal do Rio Grande do Sul, Porto Alegre, mar. 2003.

MAEDCHE, Alexander; STAAB, Steffen. Ontology Learning for the Semantic Web. **IEEE Intelligent Systems**. Arizona, v.16, n. 2. p. 72-79, Mar./Apr. 2001.

MARQUES, Manoel. C. Neto; MENDONÇA, Manoel; SANTOS, Celso Alberto Saibel. GraphMiner: Uma Ferramenta de Mineraçãõ Visual de Dados em Bases Relacionais . In: JORNADAS IBEROAMERICANAS DE INGENIERÍA DEL SOFTWARE E INGENIERÍA DEL CONOCIMIENTO - JIISIC, 4., 2004, Madrid. **Anais...** Madrid: Servicio de Publicaciones de la Facultad de Informãtica de la U.P.M, 2004, p. 305-316.

MURRAY-Rust, Peter; RZEPAB, Henry S.. STXML. A Markup Language for Scientific, Technical and Medical Publishing. **Data Science Journal**, London, v. 1, n. 2, p. 128-192, Aug. 2002.

NASCIMENTO, Hugo A. D., FERREIRA, Cristiane B. R., Visualizaçãõ de Informações: Uma Abordagem Prãtica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇãõ, 25., 2005, Sãõ Leopoldo, Brasil. **Anais...**, Sãõ Leopoldo: SBC, 2005. p. 1262-1312.

NETBEANS. **Java IDE (Integrated Development Environment)**. Disponível em: <<http://netbeans.org>>. Acesso em: Jun. 2009.

NEWSMAP. **Application for Google News**. Disponível em: <<http://newsmap.jp>>. Acesso em: Jun. 2009.

NOY, Natalya Fridman; FERGERSON, Ray W.; MUSEN, Mark A. The Knowledge model of Protégé-2000: combining interoperability and flexibility. In: INTERNATIONAL CONFERENCE IN KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT - EKAW, 12th. , 2000, Juan-Les-Pins. **Proceedings...** Berlin: Lecture Notes in Artificial Intelligence LNAI - Springer-Verlag, 2000. p 17-32.

NOY, Natalya Fridman; MCGUINNESS, Deborah. Ontology development 101: a guide to creating your first ontology. Standard: Stanford Knowledge Systems Laboratory KSL-01-05 e Stanford Medical Informatics SMI-2001-0880, Mar. 2001. Relatório Técnico.

OPENDX. **Open Visualization Data Explorer**. Disponível em: <<http://www.research.ibm.com/dx/>>. Acesso em: Jun. 2009.

OWL . **Ontology Web Language**. Disponível em: <<http://www.w3.org/2004/OWL>>. Acesso em: Mai. 2009.

OWLAPI. **API for Ontology Web Language**. Disponível em: <<http://owlapi.sourceforge.net>>. Acesso em: Mai. 2009.

PAIVA, Débora M. B.; TURINE, Marcelo A. S.; FORTES, Renata P. M. Processo de Software Livre em Ambiente Acadêmico: Experiências e Lições Aprendidas. In: CONFERÊNCIA LATINO AMERICANA DE INFORMÁTICA, 33., 2006, Santiago. **Anais...** Santiago: SBC, 2006. p. 1-10.

PORTO, Daniel. P.; MENDONÇA, Manoel.; FABBRI, Sandra. C. P. F. CRISTA - Code Reading Implemented with Stepwise Abstraction. In: SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE - SESSÃO DE FERRAMENTAS, XXII, Outubro, 2008, Campinas, Brasil. **Anais...** Campinas: SBC, 2008. p. 50-56.

POUSMAN, Zachary; STASKO, John. A taxonomy of ambient information systems: four patterns of design. In: WORKING CONFERENCE ON ADVANCED VISUAL INTERFACES - AVI, 2006, Veneza. **Proceedings...** New York: ACM Publisher, 2006, p. 67-74.

PREFUSE. **Software tools for creating rich interactive data visualizations**. Disponível em: <<http://prefuse.org>>. Acesso em: Jun. 2009.

RAO, Ramana. From research to real world with Z-GUI. **IEEE Computer Graphics and Applications**, Palo Alto, v. 17, n. 4, p. 71-73, Jul./Aug. 1997.

RAY, Erik T. Learning **XML: Creating Self-Describing**. 2ª Ed. Cambridge: O'Reilly & Associates, 2003. 416p.

RIBEIRO, Luiz Carlos Júnior. **Construção semi-automática de ontologias a partir de textos da língua portuguesa**. Dissertação (Mestrado em Computação Aplicada). Pontifícia Universidade Católica do Rio Grande do Sul, São Leopoldo, 2008.

SEMCOR. **English/Italian parallel corpus**. Disponível em: <<http://multisemcor.itc.it/index.php>>. Acesso em: Jun. 2009.

SCHREIBER, Guus; WIELINGA, Bob; JANSWEIJER, Wouter. The KACTUS view on the 'O' word. In: WORKSHOP ON BASIC ONTOLOGICAL ISSUES IN KNOWLEDGE SHARING - IJCAI, 1995, Montreal. **Proceedings...** California: AAI Press, Aug, 1995. p. 159-168.

SINAES. **Portal do Sistema Nacional de Avaliação do Ensino Superior**. Disponível em: <<http://sinaes.inep.gov.br/sinaes>>. Acesso em: mai. 2009.

SURE, York; ERDMANN, Michael; ANGELE, Juergen; STAAB, Steffen; STUDER, Rudi; WENKE, Dirk. OntoEdit: Collaborative Ontology Development for the Semantic Web. In: INTERNATIONAL SEMANTIC WEB CONFERENCE- ISWC, 2002, Sardinia. **Proceedings...** Heidelberg: Lecture Notes in Computer Science LNCS - Springer Berlin, Jun. 2002. p. 221-235.

SURYANTO, Hendra; COMPTON, Paul. Discovery of ontologies from knowledge bases. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE CAPTURE - K-CAP, 2001, Victoria. **Proceedings...** New York: ACM, 2001, p. 171-178.

SWARTOUT, Bill; PATIL, Ramesh; KNIGHT, Kevin; RUSS, Tom. Toward distributed use of large-scale ontologies. In: BANFF KNOWLEDGE ACQUISITION FOR KNOWLEDGE-BASED SYSTEMS WORKSHOP, 10th, 1996, Alberta. **Proceedings...** Calgary: SRDG Publications, 1996, p. 1 -19

TOLLIS, Ioannis G. Graph Drawing and Information Visualization. **ACM Computing Surveys**, New York, v. 28, n. 4, p. 19. Dez. 1996.

TUKEY, John.W. Exploratory Data Analysis. In: RAMSEY; WILEY. **Applied Psychological Measurement**. California: Addison-Wesley , 1977. p. 151-155.

TULIP. **Software for visualisation of huge graphs**. Disponível em: <<http://www.tulip-software.org/>>. Acesso em: Jun. 2009.

TURINE, Marcelo A. S.; CÁCERES, Edson N.; MONGELLI, Henrique; PAIVA, Débora M. B.; FABBRI, Sandra; ROCHA, Maria da Graça B.; CIFERRI, Ricardo R.; SANTOS, Marilde T. P. **Web-PIDE: uma plataforma aberta de integração e avaliação de dados educacionais**. São Carlos: DC-UFSCar: Campo Grande: DCT-UFMS. Observatório da Educação, Brasília, 2006. Projeto aprovado no Edital n.001/2006/INEP/CAPES.

USCHOLD, Mike.; KING, Martin., Towards a methodology for building ontologies. In: WORKSHOP ON BASIC ONTOLOGICAL ISSUES IN KNOWLEDGE SHARING, 1995. Montreal. **Proceedings...** Montreal: AAI Press, 1995. p. 183-196.

XCES. **Corpus Encoding Standard for XML**. Disponível em: <<http://www.xces.org/>>. Acesso em: Jun. 2009.

XMDV. **Software package for the interactive visual exploration.** Disponível em: <<http://davis.wpi.edu/xmdv/>>. Acesso em: Jun. 2009.

XML. **Extensible Markup Language (XML).** Disponível em: <<http://www.w3.org/XML/>>. Acesso em: Jun. 2009.

W3C. W3C Recommendation. Extensible Markup Language (XML) 1.0. 5th Ed.. Nov. 2008. Disponível em: <<http://www.w3.org/TR/2008/REC-xml-20081126>>. Acesso em: Jun. 2009.

WANG, Xin; LAO, Guillermo; DEMARTINI, Thomas; REDDY, Hari; NGUYEN, Mai; VALENZUELA, Edgar. XrML - eXtensible rights Markup Language. In: ACM WORKSHOP ON XML SECURITY, 2002, Fairfax. **Proceedings...** New York: ACM, 2002, p. 71-79.

WORDNET. **Lexical database of English.** Disponível em: <<http://wordnet.princeton.edu/>>. Acesso em: Junho/2009.

WOHLIN, Claes; RUNESON, Per; HÖST, Martin; OHLSSON, Magnus C.; REGNELL, Björn. & WESSLÉN, Anders. **Experimentation in software engineering - an introduction.**Sweden: Springer-Verlag, 2000. 228p.

WRIGHT, William. Business visualization applications. **IEEE Computer Graphics and Applications**, Los Alamitos, v. 17, n. 4, p. 66-7

Apêndice A

COMPARAÇÕES ENTRE ARQUIVOS DO SAEB

Esse documento foi preparado por todos os membros do projeto Web-PIDE. A falta de acentos ortográficos condiz com a realidade dos arquivos comparados. A comparação foi feita entre todos os arquivos do SAEB disponibilizados pelo INEP. Por ser um grande número de arquivos, este apêndice contém apenas as comparações dos arquivos de 2001, dos questionários de matemática.

Arquivos comparados:		Avaliação/Ano:		SAEB/2001							
pdf	LEIA-ME_SAEB_2003 (DICIONARIO_MATEMATICA_04SERIE)										
sas	INPUT_SAS_MATEMATICA_04SERIE										
Inconsistências detectadas:											
PDF	SAS	DESCRICAÇÃO PDF	DESCRICAÇÃO SAS	Início PDF	Início SAS	Tamanho PDF	Tamanho SAS	Tipo no PDF	Tipo no SAS	Observações nos Microdados	Comentários
ESTRATO	ESTRATO	Estrato da Amostra	Código do aluno	29	@29	12	12.	Numérica	Numérica		Descrição incorreta no arquivo SAS.

REG_METRO P	REG_METRO P	Localizado em Região Metropolitana?	Localizado em Região Metropolitana?	109	@109	8	8.	Numérica	Numérica	Respostas 1 e 2.	O arquivo PDF não contém a categoria e a descrição das variáveis categóricas para esta variável.
RESP_BL1	RESP_BL1	Respostas do aluno no Bloco 1 do Caderno	Respostas do aluno no Bloco 1 do Caderno	181	@181	13	\$char13.	Alfanumérica	Alfanumérica	As respostas possuem de 0 a 13 caracteres e algumas contém os caracteres ' ' (espaço) e '*' (asterisco).	
RESP_BL2	RESP_BL2	Respostas do aluno no Bloco 2 do Caderno	Respostas do aluno no Bloco 2 do Caderno	194	@194	13	\$char13.	Alfanumérica	Alfanumérica		
RESP_BL3	RESP_BL3	Respostas do aluno no Bloco 3 do Caderno	Respostas do aluno no Bloco 3 do Caderno	207	@207	13	\$char13.	Alfanumérica	Alfanumérica		
GAB_BL1	GAB_BL1	Gabarito do Bloco 1 do Caderno	Gabarito do Bloco 1 do Caderno	220	@220	13	\$char13.	Alfanumérica	Alfanumérica	Todas as respostas contém 12 caracteres.	
GAB_BL2	GAB_BL2	Gabarito do Bloco 2 do Caderno	Gabarito do Bloco 2 do Caderno	233	@233	13	\$char13.	Alfanumérica	Alfanumérica		
GAB_BL3	GAB_BL3	Gabarito do Bloco 3 do Caderno	Gabarito do Bloco 3 do Caderno	246	@246	13	\$char13.	Alfanumérica	Alfanumérica		
042M001	A042M001	Quando você começou a estudar?	Quando você começou a estudar?	750	@750	8	8.	Numérica	Numérica	Todas as respostas estão em branco.	
Exemplo da inconsistência dos microdados:											
RESP_BL1	*BADCCBBBC*DB										
RESP_BL2	CCBACBAAABDCB										
RESP_BL3	B										
GAB_BL1	CBDBDAADCADB										
GAB_BL2	CBDADACDDCA										
GAB_BL3	CADBCCDCCDD										

Arquivos comparados:		Avaliação/Ano:		SAEB/2001							
pdf	LEIA-ME_SAEB_2001 (DICIONARIO_MATEMATICA_08SERIE)										
sas	INPUT_SAS_MATEMATICA_08SERIE										
Inconsistências detectadas:											
PDF	SAS	DESCRICAÇÃO PDF	DESCRICAÇÃO SAS	Início PDF	Início SAS	Tamanho PDF	Tamanho SAS	Tipo no PDF	Tipo no SAS	Observações nos Microdados	Comentários
ANO	ANO	Ano de aplicação	Ano	17	@17	8	8.	Alfanumérica	Numérica		Tipo no PDF e no SAS não coincidem.
TURMA	TURMA	Número da turma	Numero da turma	28	@28	8	8.	Alfanumérica	Numérica		
ALUNO	ALUNO	Código do aluno	Codigo do aluno	36	@36	8	8.	Alfanumérica	Numérica		
SERIE	SERIE	Série	Serie	25	@25	2	\$Char2.	Alfanumérica	Alfanumérica	Categoria = 08.	No dicionário está Categoria = 4.
UF	uf	Unidade da Federação	Unidade da Federação	104	@104	8	8.	Numérica	Numérica	Categoria (Mato Grosso do Sul) = 50.	No dicionário está Categoria (Mato Grosso do Sul) = 5.
RESP_BL1	RESP_BL1	Respostas do aluno no Bloco 1 do Caderno	Respostas do aluno no Bloco 1 do Caderno	160	@160	13	\$Char13.	Alfanumérica	Alfanumérica	As respostas possuem de 0 a 13 caracteres e algumas contém os caracteres ' ' (espaço) e '*' (asterisco).	
RESP_BL2	RESP_BL2	Respostas do aluno no Bloco 2 do Caderno	Respostas do aluno no Bloco 2 do Caderno	173	@173	13	\$Char13.	Alfanumérica	Alfanumérica		
RESP_BL3	RESP_BL3	Respostas do aluno no Bloco 3 do Caderno	Respostas do aluno no Bloco 3 do Caderno	186	@186	13	\$Char13.	Alfanumérica	Alfanumérica		
A081_005	A081_005	Você trabalha?	Voce trabalha?	321	@321	8	8.	Numérica	Numérica	Respostas de 1 a 4.	Categoria vai de 1 a 2 no dicionário.
A081_023	A081_023	Quantas pessoas moram com voce?	Quantas pessoas moram com voce?	465	@465	8	8.	Numérica	Numérica	Respostas de 1 a 6.	Categoria vai de 1 a 5 no dicionário.
A082M033	A082M033	Da 5ª a 8ª serie, voce teve professor de Matematica	Da 5ª a 8ª serie, voce teve professor de Matematica:	881	@881	8	8.		Numérica		Está faltando o tipo no dicionário (pdf).
A082M036	A082M036	O seu professor se esforça para que você aprenda Matematica?	O seu professor se esforça para que você aprenda Matematica?	905	@905	8	8.		Numérica		Está faltando o tipo no dicionário (pdf).
Exemplo da inconsistência dos microdados:											
RESP_BL1	ed*adcaacabca										

Arquivos comparados:		Avaliação/Ano:		SAEB/2001							
pdf	LEIA-ME_SAEB_2001 (DICIONARIO_MATEMATICA_03ANO)										
sas	INPUT_SAS_MATEMATICA_03ANO										
Inconsistências detectadas:											
PDF	SAS	DESCRICAO PDF	DESCRICAO SAS	Início PDF	Início SAS	Tamanho PDF	Tamanho SAS	Tipo no PDF	Tipo no SAS	Observações nos Microdados	Comentários
ANO	ANO	Ano de aplicação	Ano	17	@17	8	8.	Alfanumérica	Numérica		Tipo no PDF e no SAS não coincidem.
TURMA	TURMA	Número da turma	Numero da turma	28	@28	8	8.	Alfanumérica	Numérica		
ALUNO	ALUNO	Código do aluno	Codigo do aluno	36	@36	8	8.	Alfanumérica	Numérica		
UPA	UPA	Unidade Primária de Amostragem	Unidade Primaria de Amostragem	60	@60	8	12.	Numérica	Numérica		Tamanho errado no dicionário (pdf).
SERIE	SERIE	Série	Serie	25	@25	2	\$Char2.	Alfanumérica	Alfanumérica	Categoria = 11.	No dicionário está Categoria = 4.
UF	uf	Unidade da Federação	Unidade da Federação	104	@104	8	8.	Numérica	Numérica	Categoria (Mato Grosso do Sul) = 50.	No dicionário está Categoria (Mato Grosso do Sul) = 5.
RESP_BL1	RESP_BL1	Respostas do aluno no Bloco 1 do Caderno	Respostas do aluno no Bloco 1 do Caderno	160	@160	13	\$Char13.	Alfanumérica	Alfanumérica	As respostas possuem de 0 a 13 caracteres e algumas contém os caracteres ' ' (espaço) e '*' (asterisco).	
RESP_BL2	RESP_BL2	Respostas do aluno no Bloco 2 do Caderno	Respostas do aluno no Bloco 2 do Caderno	173	@173	13	\$Char13.	Alfanumérica	Alfanumérica		
RESP_BL3	RESP_BL3	Respostas do aluno no Bloco 3 do Caderno	Respostas do aluno no Bloco 3 do Caderno	186	@186	13	\$Char13.	Alfanumérica	Alfanumérica		
A111_005	A111_005	Você trabalha?	Voce trabalha?	321	@321	8	8.	Numérica	Numérica	Respostas de 1 a 4.	Categoria vai de 1 a 2 no dicionário.
A111_023	A111_023	Quantas pessoas moram com voce?	Quantas pessoas moram com voce?	465	@465	8	8.	Numérica	Numérica	Respostas de 1 a 6.	Categoria vai de 1 a 5 no dicionário.
A112M027	A112M027	Da 1ª a 3ª serie, voce teve professor de Matematica	Da 1ª a 3ª serie, voce teve professor de Matematica:	849	@849	8	8.		Numérica		Está faltando o tipo no dicionário (pdf).
A112M030	A112M030	O seu professor se esforça para que você aprenda Matematica?	Voce acha que seu professor se esforça para que voce aprenda Matematica?	873	@873	8	8.		Numérica		Está faltando o tipo no dicionário (pdf).

Exemplo da inconsistência dos microdados:	
RESP_BL1	e**dbec***cda
RESP_BL2	dddddbce d
RESP_BL3	db caab

Apêndice B

OBSERVAÇÕES SOBRE OS QUESTIONÁRIOS DO SAEB

Avaliação: SAEB			Questionário: Docente	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
atividade mais relevante de que você participou nesse período	2001			Que período? É importante ter o cuidado para que o sentido das questões não dependa da ordem com que elas aparecem.
qual a atividade mais relevante de que você participou nos últimos 2 anos?	1999			Essa questão e a questão anterior querem dizer a mesma coisa?
nas reuniões sobre a gestão escolar você participa das decisões?	1997	quantas vezes você foi convidado/convocado para participar de reuniões para decisões de assuntos relacionados a gestão da escola neste ano letivo?	1995	As questões querem saber a mesma coisa?
você utiliza acesso a internet como recurso pedagógico?	1999	o acesso a internet para uso dos alunos e utilizado como recurso pedagógico?	2001	As questões querem saber a mesma coisa?
há quantos anos você está lecionando?	2003	há quantos anos você é professor(a)?	2001	As questões querem saber a mesma coisa?
há quantos anos você trabalha como professor regente de classe?	1995	há quantos anos você é professor(a)?	2001	As questões querem saber a mesma coisa?
na escolha do livro didático utilizado nesta turma, foi consultado o guia do livro didático da sef/mec?	2003	para a escolha do livro didático você considerou a relação publicada pelo mec?	1997	A relação publicada pelo MEC é o Guia do Livro Didático da SEF/MEC?
você utiliza computadores na escola?	2003	você utiliza computadores para uso dos alunos como recurso pedagógico?	1999 e 2001	A questão de 2003 refere-se à recurso pedagógico ou não?
você costuma ler revistas em quadrinhos?	2003	você leu revistas em quadrinhos neste ano?	1999	O importante é saber se o professor costuma ler ou se ele realmente leu as revistas?
você costuma ler jornais ou revistas de informação geral (veja, época)?	2003	você leu jornais ou revistas de informação geral neste ano?	1999	O importante é saber se o professor costuma ler ou se ele realmente leu as revistas de informação geral?

Avaliação: SAEB			Questionário: Turma	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Horário de início de funcionamento desta turma	1999	O turno de funcionamento desta turma	2001	Que tal padronizar essa questão em próximos questionários?
Horário de início de funcionamento desta turma	1999	Qual o horário de início da aula?	2003	
Qual o horário de final da aula?	2003	Horário de término de funcionamento desta turma	1999	
Qual o horário de início da aula?	2003	Horário de início de funcionamento desta turma	1999	

Avaliação: SAEB			Questionário: Biologia	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Como você se considera?	1997 1999			Pergunta refere-se a cor/raça. Não seria conveniente deixar explícito na questão sobre o que se trata a resposta?
Voce achou o teste de Biologia:	1997			Pergunta refere-se a grau de dificuldade (muito difícil, difícil, medio ou facil . Não seria conveniente deixar explícito na questão sobre o que se trata a resposta?
Você possui computador em casa?	1997	Quantos computadores há onde você mora?	1999	Que tal padronizar essa questão em próximos questionários?
Quantas horas por semana você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	Que tal padronizar essa questão em próximos questionários?
Sua mãe sabe ler e escrever?	1999	Qual o nível de instrução da sua mãe?	1997	Que tal padronizar essa questão em próximos questionários?
Até que série sua mãe estudou?	1999	Qual o nível de instrução da sua mãe?	1997	
Até que série seu pai estudou?	1999	Qual o nível de instrução do seu pai?	1997	Que tal padronizar essa questão em próximos questionários?
Suu pai sabe ler e escrever?	1999	Qual o nível de instrução do seu pai?	1997	
Você gosta desta disciplina?	1999	Você gosta de Biologia?	1997	É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
Qual é o mês do seu aniversário?	1999	Qual e o mês e ano do seu nascimento?	1997	Que tal padronizar essa questão em próximos questionários?
Qual é o ano do seu nascimento?	1999	Qual e o mês e ano do seu nascimento?	1997	

Quantas horas por semana você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	Que tal padronizar essa questão em próximos questionários?
Você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	
Neste ano, você teve boas notas ou conceitos nesta disciplina?	1999			É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.

Avaliação: SAEB			Questionário: Ciências	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Voce possui computador em casa?	1997	Quantos computadores há onde você mora?	1999	Que tal padronizar essa questão em próximos questionários?
Qual é o mês do seu aniversário?	1999	Qual é o mês e ano do seu nascimento?	1997	Que tal padronizar essa questão em próximos questionários?
Qual é o ano do seu nascimento?	1999	Qual é o mês e ano do seu nascimento?	1997	
Quantas horas por semana você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	Que tal padronizar essa questão em próximos questionários?
Sua mãe sabe ler e escrever?	1999	Qual o nível de instrução da sua mãe?	1997	Que tal padronizar essa questão em próximos questionários?
Até que série sua mãe estudou?	1999	Qual o nível de instrução da sua mãe?	1997	
Até que série seu pai estudou?	1999	Qual o nível de instrução do seu pai?	1997	Que tal padronizar essa questão em próximos questionários?
Suu pai sabe ler e escrever?	1999	Qual o nível de instrução do seu pai?	1997	
Você gosta de Ciências?	1997	Você gosta desta disciplina?	1999	É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
Quantas horas por semana você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	Que tal padronizar essa questão em próximos questionários?

Você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	
Neste ano, você teve boas notas ou conceitos nesta disciplina?	1999			É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
Voce lê jornais?	1997	Você leu jornal ou revista de informação geral neste ano?	1999	As questões querem saber a mesma coisa? O importante é saber se o professor lê jornais ou se ele leu jornais ou revistas no ano?

Avaliação: SAEB		Questionário: Física		
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Até que série seu pai estudou?	1999	Qual o nível de instrução do seu pai?	1997	Que tal padronizar essa questão em próximos questionários?
Até que série sua mãe estudou?	1999	Qual o nível de instrução da sua mãe?	1997	Que tal padronizar essa questão em próximos questionários?
		Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	
Como você se considera?	1997 1999			Pergunta refere-se a cor/raça. Não seria conveniente deixar explícito na questão sobre o que se trata a resposta?
Você gosta desta disciplina?	1999	Você gosta de Física?	1997	
Qual é o ano do seu nascimento?	1999	Qual é o mês e ano do seu nascimento?	1997	Que tal padronizar essa questão em próximos questionários?
Qual é o mês do seu aniversário?	1999	Qual é o mês e ano do seu nascimento?	1998	

Avaliação: SAEB		Questionário: Geografia		
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
		Você gosta desta disciplina?	1999	É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
		Você se considera:	1999	Pergunta refere-se a cor/raça. Não seria conveniente deixar explícito na questão sobre o que se trata a resposta?

Avaliação: SAEB			Questionário: História	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Neste ano, você teve boas notas ou conceitos nesta disciplina?	1999			É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
Você gosta desta disciplina?	1999	Você gosta de História?	1997	

Avaliação: SAEB			Questionário: Química	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Até que série seu pai estudou?	1999	Qual o nível de instrução do seu pai?	1997	Que tal padronizar essa questão em próximos questionários?
Até que série sua mãe estudou?	1999	Qual o nível de instrução da sua mãe?	1997	Que tal padronizar essa questão em próximos questionários?
Quantas horas por semana você trabalha?	1999	Se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997	Que tal padronizar essa questão em próximos questionários?
Você trabalha?	1999			
Você gosta desta disciplina?	1999	Você gosta de Química?	1997	É importante para a LIDE que o enunciado das questões não dependa no questionário que ela se encontra. Neste caso, é importante constar a disciplina que a questão se refere. Como muitos questionários são compartilhados, no momento de armazenar essa informação na base de dados poderia ocorrer um tratamento desses dados.
Neste ano, você teve boas notas ou conceitos nesta disciplina?	1999			
Qual é o ano do seu nascimento?	1999	Qual é o mês e ano do seu nascimento?	1997	Que tal padronizar essa questão em próximos questionários?
Qual é o mês do seu aniversário?	1999	Qual é o mês e ano do seu nascimento?	1997	

Avaliação: SAEB			Questionário: Matemática	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
Você usa o que aprendeu em Matemática na escola, no seu dia a dia?	1997 04SERIE	Aprender Matemática ajuda você a resolver os problemas do seu dia a dia?	1997 08SERIE E 03ANO	Que tal padronizar essa questão em próximos questionários?

Qual a sua idade?	2001 04SERIE	idade	2001 04SERIE, 08SERIE E 03ANO	Que tal padronizar essa questão em próximos questionários?
you teve que abandonar a escola durante algum tempo?	1995 04SERIE	you já deixou de frequentar a escola em algum período?	1997 04SERIE 08SERIE E 03ANO	Que tal padronizar essa questão em próximos questionários?
de sua opinião sobre a frase: 'eu gosto de matemática'	1995 08SERIE e 03ANO	you gosta de estudar matemática?	1995 04SERIE, 1997, 2001 e 2003	Que tal padronizar essa questão em próximos questionários?
you gosta de estudar matemática?	1995 04SERIE, 1997, 2001 e 2003	you gosta desta disciplina?	1999	Que tal padronizar essa questão em próximos questionários? É importante deixar explícito sobre qual disciplina a pergunta se refere.
you já fez ou está fazendo aula de reforço escolar?	2001	you faz algum curso fora da escola, como reforço escolar?	1999 08SERIE e 03ANO	Na primeira questão, também refere-se a cursos fora da escola? Na segunda questão da segunda coluna, refere-se somente a atividades internas da escola?
you já fez ou está fazendo aula de reforço escolar?	2001	além das aulas, you participa de atividades como aula de reforço escolar ou recuperação?	2001	
se you trabalha ou tem alguma atividade remunerada, quantas horas por dia you gasta nesse trabalho?	1997 08SERIE e 03ANO	quantas horas por semana you trabalha?	1999 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
se you trabalha ou tem alguma atividade remunerada, quantas horas por dia you gasta nesse trabalho?	1997 08SERIE e 03ANO	you trabalha?	1999 e 2001	
se you trabalha ou tem alguma atividade remunerada, quantas horas por dia you gasta nesse trabalho?	1997 08SERIE e 03ANO	you exerce alguma atividade remunerada?	1995 08SERIE e 03ANO	
		em dia de aula, quanto tempo you trabalha fora de casa?	2003	
com que frequência seus pais ou responsáveis conversam com seus professores?	2003 08SERIE e 03ANO	quantas vezes seus pais vieram a escola para conversar com professores/diretor da escola?	2001 04SERIE e 08SERIE	Que tal padronizar essa questão em próximos questionários?
com que frequência seus pais ou responsáveis conversam com o diretor da sua escola?	2003 08SERIE e 03ANO	quantas vezes seus pais vieram a escola para conversar com professores/diretor da escola?	2001 04SERIE e 08SERIE	
com que frequência seus pais ou responsáveis vão à reunião de pais na escola?	2003	quantas vezes seus pais vieram a escola para reunião de pais?	2001 04SERIE e 08SERIE	Que tal padronizar essa questão em próximos questionários?
com que frequência seus pais ou responsáveis vão à reunião de pais na escola?	2003	seu pai, mãe ou responsável participa(m) das reuniões da escola?	1997	

você faz algum curso fora da escola, como computação?	1999 08SERIE e 03ANO	você já fez ou está fazendo curso de computação?	2001	Que tal padronizar essa questão em próximos questionários?
quanto tempo você gasta para fazer as lições de casa de matemática?	1997 e 1995 04SERIE	quanto tempo por semana você costuma gastar para fazer as lições de casa de matemática?	1995 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários? O interessante é saber as horas por dia, por semana ou apenas se o aluno faz lição ou estuda em casa?
em dia de aula, quanto tempo você gasta estudando ou fazendo lições de casa?	2003 08SERIE e 03ANO			
na sua casa trabalha alguma empregada doméstica?	2003	em sua casa, trabalha empregada doméstica? quantas?	2001	Que tal padronizar essa questão em próximos questionários?
		em sua casa trabalha alguma empregada doméstica, todos os dias úteis? quantas?	1999	
você usa o que aprendeu em matemática na escola, no seu dia a dia?	1995 e 1997 04SERIE	aprender matemática ajuda você a resolver os problemas do seu dia a dia?	1997 08SERIE e 03ANO	As questões querer saber a mesma coisa?
você acha que seu professor se esforça para que você aprenda matemática?	2001 08SERIE e 03ANO 1997 08SERIE e 03ANO.	dê sua opinião sobre a frase: 'o(a) professor(a) se esforça para que eu aprenda matemática'	1995 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
leu revistas de informação geral este ano?	2001	você lê revistas de informação geral (veja, época)?	2003 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno leu ou se ele costuma ler revistas e/ou jornais? Questões sobre revistas e jornais têm necessidade de serem diferentes?
leu revistas de informação geral este ano?	2001	você leu jornal ou revista de informação geral neste ano?	1999	
existe revistas de informação geral em sua casa?	2001	na sua casa chegam revistas de informação geral (veja, época)?	2003 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
em sua casa compra-se jornal?	1995 e 1997	existe um jornal diário em sua casa?	2001	As questões querer saber a mesma coisa?
em sua casa compra-se jornal?	1995 e 1997	na sua casa chega jornal para ler?	2003	
você lê jornais?	1997 08SERIE e 03ANO e 2003	leu jornais este ano?	2001	Que tal padronizar essa questão em próximos questionários?
quantas televisões a cores há onde você mora?	1999 e 2001	na sua casa tem televisão em cores?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há televisão ou saber a quantidade, quando existir?
quantas horas por dia você assiste a programas de televisão?	1995 e 1997	em dia de aula, quanto tempo você fica vendo televisão?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber somente as horas que o aluno assiste TV em

				dias de aula ou em todos os dias da semana?
quantas pessoas moram com você?	1999 e 2001	além de você, quantas pessoas moram em sua casa?	2003	Que tal padronizar essa questão em próximos questionários?
onde você mora existe água encanada?	1999 e 2001	onde você mora chega água pela torneira?	2003	As questões não querer saber a mesma coisa?
você faz lição de casa?	1999 e 2001	você faz a lição de casa de matemática?	2003	Que tal padronizar essa questão em próximos questionários? É importante deixar explícito sobre qual disciplina a pergunta se refere, a não ser que a pergunta tenha o objetivo de saber das disciplinas em geral.
com que frequência seus pais tem conversado com seus amigos?	2001	com que frequência seus pais ou responsáveis conversam com seus amigos/ colegas de escola?	2003	As questões não querer saber a mesma coisa? A primeira questão refere-se apenas aos amigos que não são da escola?
com que frequência seus pais tem conversado com seus amigos?	2001	com que frequência seus pais ou responsáveis conversam com outros amigos seus?	2003 08SERIE e 03ANO	
quantas máquinas de lavar roupa há onde você mora?	1999 e 2001	na sua casa tem máquina de lavar roupa?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há máquina de lavar ou saber a quantidade, quando existir?
turma tem aluno bolsa escola?	2003	aluno tem bolsa escola?	2003	O questionário é destinado aos alunos. Como os alunos sabem se há alguém na turma com bolsa escola?
existe acesso a internet em sua casa?	2001	na sua casa tem computador com internet?	2003	Que tal padronizar essa questão em próximos questionários?
quantos computadores há onde você mora?	1999 e 2001	você possui computador em casa?	1997	Que tal padronizar essa questão em próximos questionários?
quantos computadores há onde você mora?	1999 e 2001	na sua casa tem computador sem internet?	2003	
você usa computador na escola?	1999	você usa o computador da escola?	1997	As questões não querer saber a mesma coisa?
você já repetiu de ano? quantas vezes?	1997, 1999 e 2001	você já foi reprovado?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas se o aluno reprovou ou quantas vezes ele reprovou?
quantas geladeiras há onde você mora?	1999 e 2001	na sua casa tem geladeira?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há geladeira ou saber a quantidade, quando existir?
você faltou as aulas neste ano?	1997 e 2001	quantos dias você faltou as aulas este ano?	1995	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno faltou ou quantos dias ele faltou?
além dos livros escolares, quantos livros há em sua casa?	2003 e 2001	quantos livros há em sua casa?	1999	Que tal padronizar essa questão em próximos questionários?
você lê livros de literatura, como romance, ficção, etc?	2003 08SERIE e 03ANO	leu livros de ficção, romances este ano?	2001	Que tal padronizar essa questão em próximos questionários?

		you leu livros neste ano?	1999	
		you vê livros de histórias infantis?	2003 04 SERIE	
you faz algum curso fora da escola, como música?	1999	you já fez ou está fazendo curso de música, teatro, dança, artes plásticas?	2001	As questões querer saber a mesma coisa?
até que série seu pai estudou?	1999 e 2003	até que série o responsável por you (como por ex seu pai, padrasto, pai de criação ou avó) estudou?	2001	Que tal padronizar essa questão em próximos questionários?
seu pai sabe ler e escrever?	1999 e 2003	qual o nível de instrução do seu pai?	1995 08SERIE e 03ANO e 1997	
até que série sua mãe estudou?	1999 e 2003	até que série a responsável por you (como por ex sua mãe, madrasta, mãe de criação ou avó) estudou?	2001	Que tal padronizar essa questão em próximos questionários?
sua mãe sabe ler e escrever?	1999 e 2003	qual o nível de instrução da sua mãe?	1995 08SERIE e 03ANO e 1997	
		até que série a pessoa indicada acima estudou?	2003	Que pessoa? É importante ter o cuidado para que o sentido das questões não dependam da ordem com que elas aparecem.
quantos aspiradores de pó há onde you mora?	1999 e 2001	na sua casa tem aspirador de pó?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há aspirador de pó ou saber a quantidade, quando existir?
qual a sua idade?	2001 04SERIE	idade	2001 04SERIE	Que tal padronizar essa questão em próximos questionários?
you já deixou de frequentar a escola em algum período?	1995 08SERIE e 03ANO, 1997 e 2003	you abandonou a escola durante algum tempo?	2001	Que tal padronizar essa questão em próximos questionários?
		you teve que abandonar a escola durante algum tempo?	1995 04SERIE	
you faz algum curso fora da escola, como línguas?	1999 08SERIE e 03ANO	you já fez ou está fazendo curso de língua estrangeira?	2001	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno está fazendo algum curso ou também se ele já fez?
quantos banheiros há onde you mora?	1999 e 2001	dentro de sua casa tem banheiro?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber se há banheiro ou quantos banheiros existem?
you consegue entender tudo que seu professor de matemática ensina?	1997	de sua opinião sobre a frase: 'eu entendo quase tudo que o professor de matemática ensina'	1995 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
you consegue entender tudo que seu professor de matemática ensina?	1997	you entende o que o(a) professor(a) ensina nas aulas de matemática?	1995 04SERIE	

você mora com seu pai ou homem responsável por você?	2001	você mora com seu pai?	1999 e 2003	Que tal padronizar essa questão em próximos questionários?
você mora com sua mãe ou mulher responsável por você?	2001	você mora com sua mãe?	1999 e 2003	Que tal padronizar essa questão em próximos questionários?
você mora:	1995	com quem você mora?	1997	No dicionário de variáveis de 1995 (arquivo leia-me.pdf), as possíveis respostas para essa pergunta são: Com pai e mãe; Somente com a mãe; Com pai ou mãe em nova união; Em outra situação. É importante deixar explícito na questão sobre o que ela se refere. Como existe em outros questionários perguntas com o mesmo objetivo, o ideal seria padronizar essa questão nos próximos questionários.
quantos freezers há onde você mora?	1999 e 2001	na sua casa tem freezer separado da geladeira?	2003	Que tal padronizar essa questão em próximos questionários?
quantos freezers há onde você mora?	1999 e 2001	na sua casa tem freezer junto a geladeira?	2003	
você leu revistas em quadrinhos neste ano?	1999 e 2001	você lê revistas em quadrinhos?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno leu ou se ele costuma ler revistas em quadrinhos?
qual é o mês e ano do seu nascimento?	1997	qual é o ano do seu nascimento?	1999, 2001 03ANO e 2003	Que tal padronizar essa questão em próximos questionários?
		data de nascimento	1995	
quantos quartos há onde você mora?	1999 e 2001	na sua casa tem quartos para dormir?	2003	Que tal padronizar essa questão em próximos questionários?
quantos rádios há onde você mora?	1999 e 2001	na sua casa tem rádio?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há rádio ou saber a quantidade, quando existir?
na sua casa tem automóvel/ carro?	2003	quantos automóveis há onde você mora?	1999 e 2001	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há automóvel ou saber a quantidade, quando existir?
quantos vídeo cassetes há onde você mora?	1999 e 2001	na sua casa tem vídeo cassete?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há vídeo cassete ou saber a quantidade, quando existir?
seu professor de matemática corrige a lição de casa na sala de aula com os alunos?	2003	seu professor de matemática corrige a lição de casa?	2003	Que tal padronizar essa questão em próximos questionários? No dicionário de variáveis de 2001 (arquivo leia-me.pdf), as possíveis respostas para essa pergunta são: Não passa lição de casa; Nunca corrige a lição de casa; Corrige lição de casa algumas vezes; Corrige a lição de casa quase sempre; Sempre corrige a lição de casa. É importante deixar explícito
o professor de matemática:	2001			

				na questão sobre o que ela se refere.
seu pai ou sua mãe conhecem algum pai, mãe de amigo seu?	1999	com que frequência seus pais tem conversado com pai/mãe/responsável de amigo seu?	2001	Que tal padronizar essa questão em próximos questionários? Em 1999 a pergunta apenas quer saber se os responsáveis conhecem, mas em 2001 a pergunta é sobre a frequência que os responsáveis do aluno conversam com responsáveis de outros alunos.
quantas horas por dia você assiste a programas de televisão?	1995 e 1997	em dia de aula, quanto tempo você fica vendo televisão?	2003	
quantas horas por dia você assiste a programas de televisão?	1995 e 1997	em dia de aula, quanto tempo você fica vendo televisão?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber somente as horas que o aluno assiste TV em dias de aula ou em todos os dias da semana?

Avaliação: SAEB			Questionário: Português	
Questão	Ano do Questionário	Questão	Ano do Questionário	Observação
você gosta de ler?	1995 04SERIE e 1997 04SERIE	dê sua opinião sobre a frase: 'eu gosto de ler'	1995 08SERIE	Que tal padronizar essa questão em próximos questionários?
você consegue entender tudo que seu professor de português ensina?	1997 08SERIE e 03ANO	dê sua opinião sobre a frase: 'eu entendo quase tudo que o professor de português ensina'	1995 03ANO	Que tal padronizar essa questão em próximos questionários?
dê sua opinião sobre a frase: 'eu gosto de português'	1995 03ANO	você gosta de estudar língua portuguesa?	1997 08SERIE e 03ANO, 2001 e 2003	Que tal padronizar essa questão em próximos questionários?
você gosta de estudar língua portuguesa?	1997 08SERIE e 03ANO, 2001 e 2003	você gosta desta disciplina?	1999	Que tal padronizar essa questão em próximos questionários? É importante deixar explícito sobre qual disciplina a pergunta se refere.
aprender português ajuda você a resolver os problemas do seu dia a dia?	1997 08SERIE e 03ANO	dê sua opinião sobre a frase: 'o português ajuda as pessoas no seu dia-a-dia'	1995 03ANO	Que tal padronizar essa questão em próximos questionários?
quanto tempo por semana você costuma gastar para fazer as lições de casa de português?	1995 08SERIE e 03ANO	quanto tempo você gasta para fazer as lições de casa de português?	1995 04SERIE e 1997	Que tal padronizar essa questão em próximos questionários? O interessante é saber as horas por dia, por semana ou apenas se o aluno faz lição ou estuda em casa?
		em dia de aula, quanto tempo você gasta estudando ou fazendo lições de casa?	2003 08SERIE e 03ANO	

você faz algum curso fora da escola, como reforço escolar?	1999 08SERIE e 03ANO	você já fez ou está fazendo aula de reforço escolar?	2001	Na primeira questão, também refere-se a cursos fora da escola? Na segunda questão da segunda coluna, refere-se somente a atividades internas da escola?
você já fez ou está fazendo aula de reforço escolar?	2001	além das aulas, você participa de atividades como aula de reforço escolar ou recuperação?	2001	
seu pai, mãe ou responsável participa(m) das reuniões da escola?	1997	com que frequência seus pais ou responsáveis vão à reunião de pais na escola?	2003	Que tal padronizar essa questão em próximos questionários?
		quantas vezes seus pais vieram a escola para reunião de pais?	2001 04SERIE e 08SERIE	
você já fez ou está fazendo curso de música, teatro, dança, artes plásticas?	2001	você faz algum curso fora da escola, como música?	1999 08SERIE e 03ANO	As questões querer saber a mesma coisa?
com que frequência seus pais ou responsáveis conversam com o diretor da sua escola?	2003 08SERIE e 03ANO	quantas vezes seus pais vieram a escola para conversar com professores/diretor da escola?	2001 04SERIE e 08SERIE	Que tal padronizar essa questão em próximos questionários?
com que frequência seus pais ou responsáveis conversam com seus professores?	2003 08SERIE e 03ANO			
aprender português ajuda você a resolver os problemas do seu dia a dia?	1997 08SERIE e 03ANO	dê sua opinião sobre a frase: 'o português ajuda as pessoas no seu dia-a-dia'	1995 03ANO	Que tal padronizar essa questão em próximos questionários?
dê sua opinião sobre a frase: 'o(a) professor(a) se esforça para que eu aprenda português'	1995 08SERIE e 03ANO	você acha que seu professor se esforça para que você aprenda língua portuguesa?	1997 08SERIE e 03ANO, 2001 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
você gosta desta disciplina?	1999	você gosta de estudar língua portuguesa?	1997 08SERIE e 03ANO, 2001 e 2003	Que tal padronizar essa questão em próximos questionários? É importante deixar explícito sobre qual disciplina a pergunta se refere.
idade	1995, 1997, 1999 e 2001	qual a sua idade?	2001 04SERIE e 2003 04SERIE	Que tal padronizar essa questão em próximos questionários?
você já fez ou está fazendo curso de computação?	2001	você faz algum curso fora da escola, como computação?	1999 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
estágios de desempenho	2003	como você considera seu desempenho em português?	1997 08SERIE e 03ANO	Para a primeira pergunta, as respostas do Dicionário de Variáveis são: Muito Crítico; Crítico; Adequado; Avançado. Essa pergunta tem alguma relação com a segunda pergunta?

com que frequência seus pais ou responsáveis conversam com outros amigos seus?	2003 08SERIE e 03ANO	com que frequência seus pais ou responsáveis conversam com seus amigos/ colegas de escola?	2003	As questões não querer saber a mesma coisa? A primeira questão refere-se apenas aos amigos que não são da escola?
		com que frequência seus pais tem conversado com seus amigos?	2001	
seu pai ou sua mãe conhecem algum pai, mãe de amigo seu?	1999	com que frequência seus pais tem conversado com pai/mãe/responsável de amigo seu?	2001	Que tal padronizar essa questão em próximos questionários? Em 1999 a pergunta apenas quer saber se os responsáveis conhecem, mas em 2001 a pergunta é sobre a frequência que os responsáveis do aluno conversam com responsáveis de outros alunos.
até que série seu pai estudou?	1999 e 2003	até que série o responsável por você (como por ex seu pai, padrasto, pai de criação ou avô) estudou?	2001	Que tal padronizar essa questão em próximos questionários?
seu pai sabe ler e escrever?	1999 e 2003	qual o nível de instrução do seu pai?	1995 08SERIE e 03ANO e 1997	
até que série sua mãe estudou?	1999 e 2003	até que série a responsável por você (como por ex sua mãe, madrasta, mãe de criação ou avó) estudou?	2001	Que tal padronizar essa questão em próximos questionários?
sua mãe sabe ler e escrever?	1999 e 2003	qual o nível de instrução da sua mãe?	1995 08SERIE e 03ANO e 1997	
		até que série a pessoa indicada acima estudou?	2003	Que pessoa? É importante ter o cuidado para que o sentido das questões não dependam da ordem com que elas aparecem.
leu revistas de informação geral este ano?	2001	você lê revistas de informação geral (veja, época)?	2003 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno leu ou se ele costuma ler revistas e/ou jornais? Questões sobre revistas e jornais têm necessidade de serem diferentes?
leu revistas de informação geral este ano?	2001	você leu jornal ou revista de informação geral neste ano?	1999	
existe revistas de informação geral em sua casa?	2001	na sua casa chegam revistas de informação geral (veja, época)?	2003 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
em sua casa compra-se jornal?	1995 e 1997	existe um jornal diário em sua casa?	2001	As questões querer saber a mesma coisa?
em sua casa compra-se jornal?	1995 e 1997	na sua casa chega jornal para ler?	2003	
você lê jornais?	1997 08SERIE e 03ANO e 2003	leu jornais este ano?	2001	Que tal padronizar essa questão em próximos questionários?
além dos livros escolares, quantos livros há em sua casa?	2003 e 2001	quantos livros há em sua casa?	1999	Que tal padronizar essa questão em próximos questionários?

você lê livros de literatura, como romance, ficção, etc?	2003 08SERIE e 03ANO	leu livros de ficção, romances este ano?	2001	Que tal padronizar essa questão em próximos questionários?
		você leu livros neste ano?	1999	
		você vê livros de histórias infantis?	2003 04 SERIE	
você faz algum curso fora da escola, como línguas?	1999 08SERIE e 03ANO	você já fez ou está fazendo curso de língua estrangeira?	2001	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno está fazendo algum curso ou também se ele já fez?
você mora com seu pai ou homem responsável por você?	2001	você mora com seu pai?	1999 e 2003	Que tal padronizar essa questão em próximos questionários?
você mora com sua mãe ou mulher responsável por você?	2001	você mora com sua mãe?	1999 e 2003	Que tal padronizar essa questão em próximos questionários?
você mora:	1995	com quem você mora?	1997	No dicionário de variáveis de 1995 (arquivo leia-me.pdf), as possíveis respostas para essa pergunta são: Com pai e mae; Somente com a mae; Com pai ou mae em nova uniao; Em outra situacao. É importante deixar explícito na questão sobre o que ela se refere. Como existe em outros questionários perguntas com o mesmo objetivo, o ideal seria padronizar essa questão nos próximos questionários.
onde você mora existe água encanada?	1999 e 2001	onde você mora chega água pela torneira?	2003	As questões não querer saber a mesma coisa?
na sua casa tem televisão a cores?	2003	quantas televisões a cores há onde você mora?	1999 e 2001	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há televisão ou saber a quantidade, quando existir?
quantas horas por dia você assiste a programas de televisão?	1995 e 1997	em dia de aula, quanto tempo você fica vendo televisão?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber somente as horas que o aluno assiste TV em dias de aula ou em todos os dias da semana?
existe acesso a internet em sua casa?	2001	na sua casa tem computador com internet?	2003	Que tal padronizar essa questão em próximos questionários?
quantos computadores há onde você mora?	1999 e 2001	você possui computador em casa?	1997	Que tal padronizar essa questão em próximos questionários?
quantos computadores há onde você mora?	1999 e 2001	na sua casa tem computador sem internet?	2003	
você usa computador na escola?	1999	você usa o computador da escola?	1997	As questões não querer saber a mesma coisa?
na sua casa trabalha alguma empregada doméstica?	2003	em sua casa, trabalha empregada doméstica? quantas?	2001	Que tal padronizar essa questão em próximos questionários?
		em sua casa trabalha alguma empregada doméstica, todos os dias úteis? quantas?	1999	

qual é o mês e ano do seu nascimento?	1997	qual é o ano do seu nascimento?	1999, 2001 03ANO e 2003	Que tal padronizar essa questão em próximos questionários?
		data de nascimento	1995	
quantas máquinas de lavar roupa há onde você mora?	1999 e 2001	na sua casa tem máquina de lavar roupa?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há máquina de lavar ou saber a quantidade, quando existir?
quantos quartos há onde você mora?	1999 e 2001	na sua casa tem quartos para dormir?	2003	Que tal padronizar essa questão em próximos questionários?
você faltou as aulas neste ano?	1997 e 2001	quantos dias você faltou as aulas este ano?	1995	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno faltou ou quantos dias ele faltou?
você mora com sua mãe ou mulher responsável por você?	2001	você mora com sua mãe?	1999 e 2003	Que tal padronizar essa questão em próximos questionários?
seu professor de língua portuguesa corrige a lição de casa na sala de aula com os alunos?	2003	seu professor de língua portuguesa corrige a lição de casa?	2003	Que tal padronizar essa questão em próximos questionários? No dicionário de variáveis de 2001 (arquivo leia-me.pdf), as possíveis respostas para essa pergunta são: Não passa lição de casa; Nunca corrige a lição de casa; Corrige lição de casa algumas vezes; Corrige a lição de casa quase sempre; Sempre corrige a lição de casa. É importante deixar explícito na questão sobre o que ela se refere.
o professor de língua portuguesa:	2001			
você já repetiu de ano? quantas vezes?	1997, 1999 e 2001	você já foi reprovado?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas se o aluno reprovou ou quantas vezes ele reprovou?
quantas geladeiras há onde você mora?	1999 e 2001	na sua casa tem geladeira?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há geladeira ou saber a quantidade, quando existir?
quantos freezers há onde você mora?	1999 e 2001	na sua casa tem freezer separado da geladeira?	2003	Que tal padronizar essa questão em próximos questionários?
quantos freezers há onde você mora?	1999 e 2001	na sua casa tem freezer junto a geladeira?	2003	
se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997 08SERIE e 03ANO	quantas horas por semana você trabalha?	1999 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários?
se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997 08SERIE e 03ANO	você trabalha?	1999 e 2001	
se você trabalha ou tem alguma atividade remunerada, quantas horas por dia você gasta nesse trabalho?	1997 08SERIE e 03ANO	você exerce alguma atividade remunerada?	1995 08SERIE e 03ANO	

		em dia de aula, quanto tempo você trabalha fora de casa?	2003	
na sua casa tem automóvel/ carro?	2003	quantos automóveis há onde você mora?	1999 e 2001	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há automóvel ou saber a quantidade, quando existir?
quantos rádios há onde você mora?	1999 e 2001	na sua casa tem rádio?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há rádio ou saber a quantidade, quando existir?
quantos vídeo cassetes há onde você mora?	1999 e 2001	na sua casa tem vídeo cassete?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há vídeo cassete ou saber a quantidade, quando existir?
turma tem aluno bolsa escola?	2003	aluno tem bolsa escola?	2003	O questionário é destinado aos alunos. Como os alunos sabem se há alguém na turma com bolsa escola?
quantos aspiradores de pó há onde você mora?	1999 e 2001	na sua casa tem aspirador de pó?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber apenas de há aspirador de pó ou saber a quantidade, quando existir?
você leu revistas em quadrinhos neste ano?	1999 e 2001	você lê revistas em quadrinhos?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber se o aluno leu ou se ele costuma ler revistas em quadrinhos?
você faz lição de casa?	1999 e 2001	você faz a lição de casa de língua portuguesa?	2003	Que tal padronizar essa questão em próximos questionários? É importante deixar explícito sobre qual disciplina a pergunta se refere, a não ser que a pergunta tenha o objetivo de saber das disciplinas em geral.
quantas pessoas moram com você?	1999 e 2001	além de você, quantas pessoas moram em sua casa?	2003	Que tal padronizar essa questão em próximos questionários?
quantos banheiros há onde você mora?	1999 e 2001	dentro de sua casa tem banheiro?	2003	Que tal padronizar essa questão em próximos questionários? O importante é saber se há banheiro ou quantos banheiros existem?
quanto tempo você gasta para fazer as lições de casa de português?	1997 e 1995 04SERIE	quanto tempo por semana você costuma gastar para fazer as lições de casa de português?	1995 08SERIE e 03ANO	Que tal padronizar essa questão em próximos questionários? O interessante é saber as horas por dia, por semana ou apenas se o aluno faz lição ou estuda em casa?
em dia de aula, quanto tempo você gasta estudando ou fazendo lições de casa?	2003 08SERIE e 03ANO			