

## **Instrumentos de informação para extração de indicadores e de informação textual**

Tatiana S. Gonzaga (Grupo Stela/UFSC) [tatianag@stela.ufsc.br](mailto:tatianag@stela.ufsc.br)

Fabiano D. Beppler (Grupo Stela/UFSC) [fbeppler@stela.ufsc.br](mailto:fbeppler@stela.ufsc.br)

Aran B. T. Morales (Grupo Stela/UFSC) [aran@stela.ufsc.br](mailto:aran@stela.ufsc.br)

Rodolfo Almeida (Grupo Stela/UFSC) [ralmeida@stela.ufsc.br](mailto:ralmeida@stela.ufsc.br)

### **Resumo**

*Entre os Sistemas de Informação utilizados no auxílio aos processos decisórios de muitas organizações, é possível citar as técnicas de Data Warehousing e OLAP para o armazenamento e a recuperação de informação. A partir do estudo e da aplicação dessas tecnologias, este trabalho propõe instrumentos para extração de indicadores e informação textual sobre o Data Warehouse, o que possibilita análises multidimensionais, recuperação textual e a utilização integrada de ambos, isto é, a visualização de informação textual a partir de indicadores. Como aplicação para a validação do modelo, é apresentado o estudo de caso do Diretório de Grupos de Pesquisa no Brasil.*

*Palavras chave: Sistema de Informação, Data Warehouse, Apoio à decisão.*

### **1. Introdução**

A informação é um insumo básico nas organizações de hoje como elemento de apoio às decisões gerenciais. Muitas organizações possuem projetos de informática, representados pelos seus Sistemas Transacionais, i.e., sistemas de apoio às operações do dia-a-dia. A existência da informação operacional não garante sua disponibilidade para suprir as necessidades gerenciais. De outra forma, pode-se afirmar que a informação está disponível, mas a um custo de processamento muito elevado, o que pode comprometer o funcionamento dos Sistemas Transacionais por motivos de desempenho. Além disso, a interpretação e a extração de informações gerenciais dependem de recursos humanos especializados, o que dificulta o acesso às informações por parte dos tomadores de decisão das organizações.

Uma abordagem usual para a construção de Sistemas de Informação gerencial é o desenvolvimento de repositórios especializados conhecidos como *Data Warehouse* (DW) (KIMBALL & ROSS, 2002.). Os DWs permitem às organizações o acesso à informação gerencial de forma rápida e eficaz, evitando os problemas de confiabilidade nas informações fornecidas e de penalização do desempenho dos Sistemas Transacionais. Os projetos de informação incorporam, também, ferramentas de processamento analítico da informação (OLAP – *On-line Analytical Processing*) para os usuários de negócio.

Baseado nos conceitos de Sistemas de Informação, DW e OLAP, o presente trabalho apresenta um conjunto de instrumentos para extração de indicadores e de informação textual a partir de bases de dados sobre ciência e tecnologia. O primeiro instrumento é um modelo de DW que comporta as informações a serem utilizadas para extração. Os outros instrumentos – Plano Tabular e Busca Textual – fazem o acesso e a apresentação da informação do DW ao usuário final. Para validar esses instrumentos de informação, foram utilizadas as informações referentes ao Diretório de Grupos de Pesquisa no Brasil (DGP) da Plataforma Lattes, projeto mantido pelo CNPq desde 1992.

A seção seguinte explora os conceitos de Sistemas de Informação, DW e OLAP. Na seção 3, apresenta-se o Diretório de Grupos de Pesquisa e sua estrutura de informação operacional. A

seção 4 descreve os instrumentos construídos para extração de indicadores e informação textual. Na seção 5 são apresentadas as conclusões do trabalho, discutindo-se as implicações para a gestão da informação e do conhecimento.

## 1. Sistemas de Informação, *Data Warehouse* e OLAP

Os Sistemas de Informação (SI) são apresentados como soluções para ajudar a organização a lidar com os dados e as informações, para que esses dados possam auxiliar o dia-a-dia da organização e serem utilizados para os processos de tomada de decisão. Laudon e Laudon (1998) definem Sistema de Informação como um conjunto de componentes inter-relacionados que trabalham para coletar, processar, armazenar e distribuir informação com a finalidade de facilitar o planejamento, o controle, a coordenação, a análise e o processo decisório nas organizações.

Segundo O'Brian (2001), os Sistemas de Informação desempenham três papéis vitais em qualquer tipo de organização:

- Suporte de seus processos e suas operações;
- Suporte nas tomadas de decisão de seus funcionários e gerentes;
- Suporte em suas estratégias em busca de vantagem competitiva.

Ao processo de preparar os dados de um Sistema de Informação operacional de forma a ter uma fonte de informações que possam dar suporte ao processo de tomada de decisões deu-se o nome de *Data Warehousing*. Este processo proporciona uma sólida e concisa integração dos dados da organização para a realização de análises gerenciais estratégicas. Inmon (2002) define *Data Warehouse* como um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo de apoio às decisões gerenciais.

De acordo com Kimball e Ross (2002), os principais objetivos de um DW são: (a) permitir fácil acesso à informação; (b) apresentar informações consistentes e confiáveis; (c) ser adaptável a mudanças; (d) garantir a segurança no que se refere ao acesso às informações; (e) armazenar dados que servirão de base para o processo de tomada de decisões; e (f) ser aceitável pela comunidade de usuários.

A modelagem dimensional tem sido o mais adequado na implementação de um projeto de DW. Ela é uma técnica de projeto lógico que busca apresentar os dados em uma estrutura padronizada, intuitiva, que permite alto desempenho de acesso, específico para suportar processamento analítico. Um dos tipos de estrutura ou esquema utilizado num DW é o esquema estrela (*star schema*), o qual é composto de dois tipos de tabelas – Fato e Dimensão. As tabelas de Fato contêm as medições do negócio, já as tabelas de Dimensão armazenam os dados descritivos do negócio (KIMBALL & ROSS, 2002; TODESCO; SOUZA & MARTINS, 2002).

Para se ter acesso aos dados de um DW, há necessidade de ferramentas específicas para extração de informação. Uma dessas ferramentas é chamada de OLAP, utilizada para estruturas de múltiplas dimensões. Segundo Kimball et al. (1998), OLAP constitui-se de todas as atividades gerais e específicas de consulta e apresentação de dados numéricos, e textos provenientes do DW.

Thomsen (2002) explica que para se utilizar OLAP sobre um DW é necessário que alguns requisitos funcionais sejam preenchidos, como: (a) estruturação dimensional com referência hierárquica; (b) especificação eficaz das dimensões e cálculos dimensionais; (c) separação entre estrutura e representação; (d) flexibilidade; (e) velocidade para dar suporte à análise

ocasional; e (f) suporte multiusuário. O presente relato sumaria os requisitos e os resultados de um projeto de DW sobre o Diretório de Grupos de Pesquisa no Brasil, descrito a seguir.

## **2. Plataforma Lattes: o Diretório de Grupos de Pesquisa no Brasil**

O Diretório de Grupos de Pesquisa no Brasil (DGP), projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), contém uma base de dados mantida desde 1992. Originou-se em 1991, a partir de uma proposta de elaboração de um Almanaque de Pesquisa no CNPq bem como de um levantamento de grupos de pesquisa realizado pelo Fórum Nacional de Pró-Reitores de Pesquisa. O intuito era organizar o Programa de Laboratórios Associados encomendado em 1990 pela então Secretaria de Ciência e Tecnologia (atual Ministério de Ciência e Tecnologia) (GUIMARÃES, 1994). O objetivo do projeto era oferecer um suporte informacional atualizado sobre as atividades de pesquisa através de bases censitárias sobre todos os grupos de pesquisa em atividade no País.

Hoje, o DGP tem o claro objetivo de ser uma plataforma de informação sobre o parque científico e tecnológico brasileiro (CNPq, 2002). O esforço empreendido pelo Brasil após a Segunda Grande Guerra gerou o maior parque de C&T da América Latina. Entretanto, ainda há carência de informação organizada a respeito, o que enfraquece e dificulta a tomada de decisão sobre os desígnios da C&T nacional. Tal fato transforma o DGP em um instrumento essencial para a gestão de C&T (MARTINS & GALVÃO, 1994).

O DGP possui três importantes finalidades, que são: (a) fortalecer o intercâmbio entre pesquisadores brasileiros bem como entre estes e os pesquisadores estrangeiros; (b) preservar a memória da atividade de pesquisa; e (c) ser estratégica para as atividades de planejamento do CNPq. Esta terceira finalidade é de vital importância em processos de avaliação e acompanhamento (A&A). Por sua vez, os procedimentos de A&A são fundamentais para a tomada de decisão no âmbito do CNPq, quer em nível estratégico, quer no âmbito gerencial, como, por exemplo, na formulação de políticas de investimentos em C&T (CNPq, 2002).

Desde a versão 3.0, de 1997, o DGP “é capaz de descrever com precisão os limites e o perfil geral da atividade científico-tecnológica no Brasil” (CNPq, 2002). Igualmente, é capaz de fornecer aos interessados uma grande e diversificada massa de informação sobre detalhes de quem realiza as atividades, como e onde se realizam e sobre o quê. Desde a versão 4.0, de 2000, o DGP está integrado à Plataforma Lattes (conjunto de sistemas computacionais do CNPq que visa compatibilizar e integrar as informações em toda interação da Agência com seus usuários) por meio do Sistema de Currículos Lattes (instrumento de captura de informações curriculares integrante da Plataforma Lattes). Neste ano de 2004, está sendo realizado o sexto Censo do Diretório de Grupos de Pesquisa.

As informações constantes na base de dados do DGP incluem todos os elementos necessários para mapear a pesquisa científica e tecnológica nacional a partir da unidade de grupo de pesquisa. Um grupo de pesquisa é definido como um conjunto de indivíduos organizados hierarquicamente em torno de uma ou, eventualmente, duas lideranças, com as seguintes características (CNPq, 2002):

- O fundamento organizador dessa hierarquia é a experiência, o destaque e a liderança no terreno científico ou tecnológico;
- Existe envolvimento profissional e permanente com a atividade de pesquisa;
- O trabalho se organiza em torno de linhas comuns de pesquisa;
- Compartilham-se, em algum grau, instalações e equipamentos.

Em função dessa definição, o DGP contém informações sobre recursos humanos atuantes nos grupos (pesquisadores e estudantes), as linhas de pesquisa em andamento, as especialidades do conhecimento, os setores de aplicação/atividade envolvidos e as interações dos grupos com o setor produtivo (CNPq, 2002). Para armazenar essa informação em uma base de dados, são necessárias 76 tabelas com uma complexa malha de interligações, apenas para o esquema do banco de dados operacional. Para utilizar essa base de dados na extração de informações gerenciais, a complexidade da estrutura de informações torna a consulta custosa e com tendência à lentidão.

O alto grau de conexão entre os assuntos descritos no esquema do banco de dados operacional o faz inadequado à recuperação de informação gerencial. Há necessidade de um repositório que viabilize esse processamento. Por outro lado, as análises realizadas do DGP não dependem somente dos dados que constam na base de dados operacional. Através do Sistema de Currículo Lattes, são realizados, também, cruzamentos com os currículos dos pesquisadores e estudantes que participam dos grupos de pesquisa, com o fomento desses recursos humanos (Fomento-CNPq – base de dados operacional do CNPq que armazena informações sobre o fomento da Agência) e com informações provenientes das bases de dados da Capes (Data-Capes – instrumento de coleta de informações da Capes sobre a pós-graduação nacional). Esse é o cenário que deu origem ao desenvolvimento de um DW para atender às necessidades de informação gerencial, descrito a seguir.

### 3. Desenvolvimento e resultados

Para resolver o problema de complementação e extração de informações do DGP, foi construído um conjunto de instrumentos de informação para a extração de indicadores e informação textual.

O primeiro instrumento foi a definição de um DW que permitiu a complementação das informações operacionais do DGP e sua modelagem por assunto de negócio. Sobre o DW, foi possível a construção de novos instrumentos de análise e recuperação textual das informações. A metodologia seguida para a construção do DW e dos instrumentos citados seguiu três passos: (1) divisão do conteúdo por assuntos de negócio; (2) compreensão da forma como os assuntos podem ser explorados, isto é, quais são as visões através das quais podemos abordar os assuntos; e (3) verificação das informações relevantes de cada assunto a serem apresentadas ao usuário.

O Quadro 1 sumaria os assuntos, as visões de cada assunto e as informações a serem disponibilizadas para cada assunto.

Assuntos	Visões	Informações
Grupos de pesquisa	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área; Ano de formação.	Total de grupos de pesquisa Total de linhas de pesquisa Total de pesquisadores Total de estudantes Total de técnicos
Linhas de pesquisa	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área, Setor de aplicação.	Total de linhas de pesquisa Total de pesquisadores Total de grupos de pesquisa
Pesquisadores	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área, Setor de aplicação; Faixa etária.	Total por titulação máxima Total por sexo Total de líderes Total por nacionalidade

Estudantes	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área; Faixa etária.	Total por nível de treinamento Total por sexo Total por nacionalidade
Técnicos	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área; Atividade técnica.	Total geral Total com 1º e 2º graus Total com 3º grau Total com mestrado Total com doutorado
Produção CT&A	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área; Ano da produção.	Total por tipo de produção Total por tipo de autor: (pesquisadores, estudantes, pesquisadores doutores)
Grupos/ Empresas	Brasil, Região, Unidade da Federação, Instituição; Grande área, Área.	Total de grupos Total de empresas Total por tipo de relacionamento Total por tipo de remuneração
Empresas/ Grupos	Brasil, Região, Unidade da Federação; Ramo de atividade, Natureza jurídica, Pessoal ocupado.	Total de empresas Total de grupos Total por tipo de relacionamento Total por tipo de remuneração

Quadro 1 - Assuntos, visões e informações do Diretório de Grupos de Pesquisa

A partir dessas necessidades de informação, elaborou-se o modelo do DW. A Figura 1 ilustra um fragmento desse modelo – uma tabela de Fato e suas respectivas Dimensões para o assunto grupo de pesquisa modelado no DW. A tabela de Fato FT\_GRUPO armazena as medidas relacionadas às informações referentes ao assunto grupo de pesquisa. As dimensões descrevem as informações relacionadas a grupos de pesquisa.

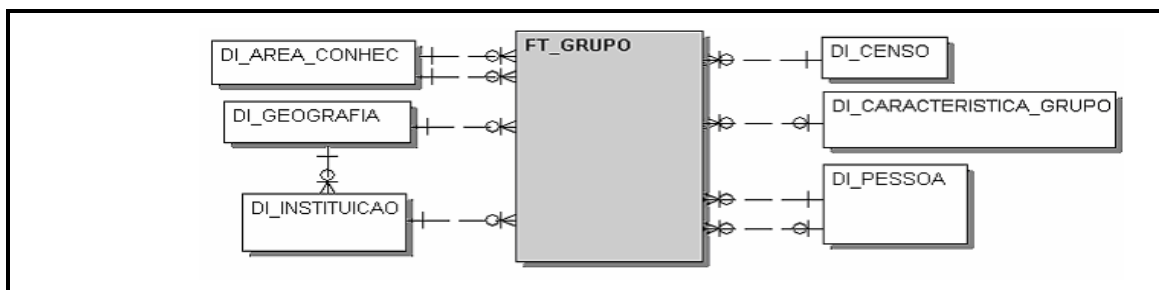


Figura 1 - Exemplo de tabela de Fato e Dimensões do DW do Diretório de Grupos de Pesquisa

Até o segundo Censo, as informações do DGP eram enviadas ao CNPq em papel, e os indicadores eram construídos individualmente, de forma manual. A partir do terceiro Censo, em 1997, a submissão das informações pelos grupos passou a ser digital, embora os indicadores ainda fossem gerados manualmente. A partir da versão 4.0, em 2000, os indicadores do DGP passaram a ser baseados no DW aqui relatado, que vem sendo aperfeiçoado a cada nova realização do censo bianual.

O Quadro 2 é um comparativo das características de construção e utilização do DGP em dois momentos – até a versão 3.0, com construção manual de indicadores a partir de informações operacionais, e da versão 4.0 em diante, com a construção de indicadores baseada no DW.

Característica de construção ou utilização	Utilizando a base operacional (v. 3.0)	Utilizando a base DW (v. 4.0, 5.0 e 6.0)
Forma de extração de informações	Consultas predefinidas, acessadas em páginas <i>Web</i> estáticas.	Consultas configuráveis pelo usuário, acessadas em páginas <i>Web</i> dinâmicas.
Cruzamento entre informações	Conjunto limitado de indicadores previamente definidos.	O limite é a quantidade de cruzamentos possíveis entre todas as variáveis do DW. O usuário define o tipo de cruzamento entre as variáveis.
Tipos de informações	Limite de informações em função do tempo gasto para geração de cada indicador.	O DW compreende o conjunto total de informações sobre grupos de pesquisa, além de conter informações sobre a produção dos integrantes dos grupos.
Construção das consultas	Demorada, pois é necessária a intervenção de um especialista humano que gera individualmente as consultas a serem disponibilizadas.	Rápida, pois há sistemas que geram indicadores automaticamente a partir do DW, não sendo necessário enfoque individual para cada consulta.
Quantidade de consultas	Limitada somente às consultas que se projetou disponibilizar. Por exemplo, a unidade Grupo de Pesquisa disponibiliza somente 172 consultas.	Limitada à quantidade de cruzamentos possíveis disponibilizados pelo DW. Por exemplo, a unidade Grupo de Pesquisa disponibiliza 13.699 consultas, que representam as possibilidades de cruzamento entre as variáveis relacionadas a grupos de pesquisa.
Alcance das consultas	Apenas os cruzamentos disponibilizados em páginas estáticas.	Além de todos os cruzamentos possíveis, qualquer conteúdo pode ser encontrado através da Busca Textual.
Integração entre ferramentas	Não há integração entre quaisquer das ferramentas.	Há integração entre extração de indicadores e busca textual, pois o DW compreende os dados para geração de indicadores e para busca textual.
Tempo gasto para a disponibilização dos resultados aos usuários	1 ano (tempo necessário para a análise individual de cada consulta e a construção das páginas estáticas).	3 meses (tempo necessário para construção do DW e das ferramentas para apresentação dos resultados).

Quadro 2 - Comparação de atividades executadas para apresentação de dados no DGP

O segundo instrumento é o Plano Tabular, um módulo de extração automática de indicadores, definido como uma aplicação OLAP. Esse instrumento é responsável por gerar consultas dinâmicas, baseadas na definição de escolhas de variáveis configuradas pelo usuário, e por apresentar os resultados da extração dos indicadores. Conforme relata o Quadro 2, e como o DW armazena informações de forma hierarquizada, tornou-se possível a recuperação de informações de maneira hierarquizada. Também é possível fazer todos os cruzamentos dos assuntos e de suas correspondentes visões.

O terceiro instrumento é a Busca Textual, uma ferramenta que proporciona a busca configurada e a visualização textual das informações sobre os grupos de pesquisa. Como o DW prevê o armazenamento das informações textuais, o instrumento de extração de indicadores permite buscar e visualizar, de forma textual, o detalhamento de cada elemento da tabela resultante. No DW, a informação textual do grupo de pesquisa é armazenada em forma de texto usando a linguagem de marcação XML. Esse armazenamento das informações textuais no DW proporciona a disponibilização de filtros que utilizam as demais informações

contidas no DW, agilizando também o processo de recuperação de informações em dois aspectos: (a) em termos de velocidade, pois a estrutura está projetada para privilegiar o desempenho e não a eliminação de redundância – como é o caso em bases operacionais; e (b) em termos de facilidade de extração, pois o DW está organizado de tal forma a facilitar a montagem dinâmica de consultas ao DW.

Outra vantagem de se ter armazenado no mesmo DW a informação textual e os dados gerenciais é a possibilidade de, a partir do resultado de uma consulta no Plano Tabular, visualizar o detalhamento textual de cada um dos elementos resultantes da consulta, sendo cada indicador visualizado detalhadamente de forma textual. Dessa forma, é possível unir a vantagem de utilização de um DW com a recuperação textual da informação. Analisando essa forma de visualização detalhada de indicadores, é possível afirmar que esse modelo permite a busca textual de informação a partir da recuperação de indicadores. Esta é uma das novas características do DGP em sua versão 6.0, de 2004.

Na Figura 2 são apresentados *screenshots* do Plano Tabular e Busca Textual, os instrumentos de visualização de informações gerenciais resultantes deste trabalho. Ambos estão disponíveis à comunidade no *web site* do CNPq, especificamente no portal da Plataforma Lattes (<http://lattes.cnpq.br>).

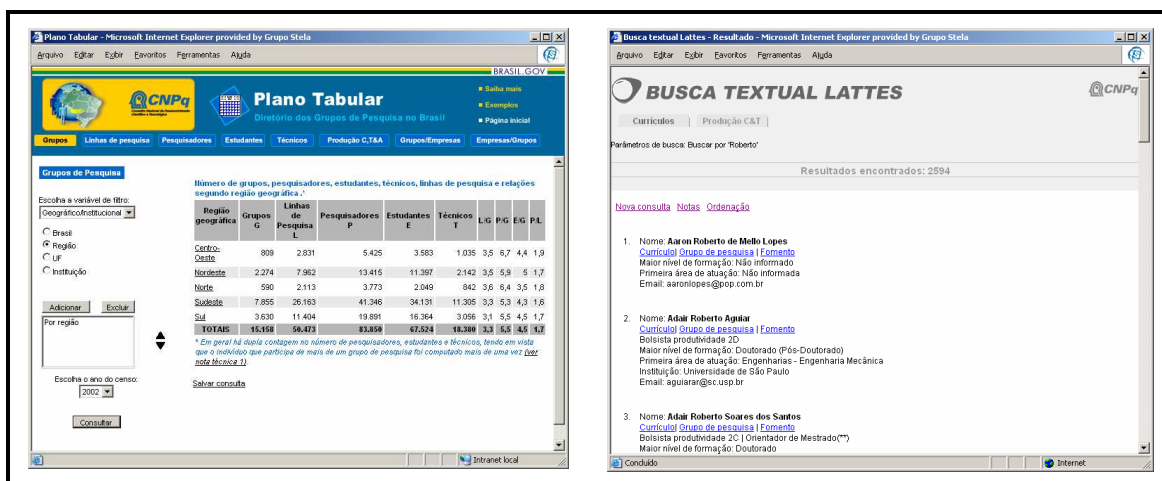


Figura 2 - Visualização das ferramentas Plano Tabular e Busca Textual

#### 4. Conclusão

Este trabalho apresentou um conjunto de instrumentos para a extração de informações gerenciais. Foi evidenciada a importância da construção do DW como repositório das informações a serem analisadas com a modelagem orientada ao assunto do negócio das análises, em comparação à modelagem segundo o processo do negócio, que é a lógica de construção de Sistemas de Informação operacional.

A disponibilidade do DW agiliza a construção de instrumentos para a extração de indicadores e informação textual, bem como agiliza o processamento desses instrumentos. Devido à estrutura de informações contidas num DW, além de disponibilizar instrumentos mais amigáveis para os usuários, o DW ainda possibilita a extração de informações com maior velocidade.

Com respeito ao DGP, na construção de indicadores para a tomada de decisão operacional e gerencial, pode-se destacar que a implementação do DW tem o impacto descrito a seguir:

- Integração das informações do DGP com as informações de outras bases de dados (Currículos Lattes dos integrantes – pesquisadores e estudantes, DW Fomento do CNPq e Data-Capes). Essa integração favorece a qualidade da informação, uma vez que se eliminam dados redundantes e possivelmente inconsistentes, bem como favorece a oportunidade de criar novas funções, eliminar trabalho sem valor agregado e construir novas informações e conhecimentos – por exemplo, “na área de conhecimento ‘Física’, na última década, qual foi a contabilização de recursos recebidos por cada grupo de pesquisa, seus índices de publicação em revistas indexadas e seus índices de formação de mestres e doutores?”.
- Disponibilização de uma ferramenta de extração de indicadores de grupos de pesquisa, o Plano Tabular que, além de oferecer ao usuário final um sistema de consulta amigável, não técnico e eficiente, oferece milhares cruzamentos de informações sobre os vários assuntos presentes na base do DGP.
- Visualização de informações textuais dos grupos de pesquisa a partir da extração de indicadores, através do sistema de Busca Textual. Essas informações englobam, por exemplo: o nome do grupo de pesquisa, sua linha de pesquisa, seu líder e integrantes (pesquisadores, estudantes) e a produção dos integrantes dos grupos.
- Rápida disponibilização dos instrumentos de informação para a comunidade a partir do término da coleta de dados em cada censo da pesquisa brasileira.
- Apresentação de forma clara e objetiva das informações sobre a pesquisa nacional.
- Auxílio ao processo de tomada de decisão em relação a C&T no País.

Na linha de futuros desenvolvimentos, há a possibilidade de utilizar o DW para extração de conhecimento através de técnicas de *data mining*, como, por exemplo, algoritmos de link análises para redes de pesquisa, que, entre outras coisas, pode verificar o tipo da relação entre integrantes de diferentes grupos.

## 5. Referências bibliográficas

- CNPq. (2002) - *Diretório de Grupos de Pesquisa no Brasil – Censo 2002*. Conselho Nacional de Desenvolvimento Científico e Tecnológico. Brasília. Disponível em: <<http://lattes.cnpq.br/censo2002/>>.
- GUIMARÃES, R. (1994) - *Avaliação e Fomento de C&T no Brasil: propostas para os anos 90*. MCT/CNPq. Brasília.
- INMON, W. H. (2002) - *Building the Data Warehouse*. John Wiley & Sons. 3<sup>rd</sup> ed. New York.
- KIMBALL, R. et al. (1998) - *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses*. John Wiley & Sons. New York.
- KIMBALL, R. & ROSS, M. (2002) - *The Data Warehouse Toolkit: The complete guide to dimensional modeling*. John Wiley & Sons. 2<sup>a</sup> Edição. New York.
- LAUDON, K. C. & LAUDON, J. P. (1998) - *Management Information Systems: New Approaches to Organization & Technology*. Prentice Hall. New Jersey.
- MARTINS, G. M. & GALVÃO, G. (1994) - *Diretório de Grupos de Pesquisa no Brasil: Perspectivas de Fomento e Avaliação. Educação Brasileira*. Vol. 16, n. 33, p. 11-29.
- O'BRIAN, J. A. (2001) - *Sistemas de Informação e as Decisões Gerenciais na Era da Internet*. Saraiva. 9<sup>a</sup> São Paulo.
- THOMSEN, E. (2002) - *OLAP – Construindo Sistemas de Informações Multidimensionais*. Campus. Rio de Janeiro.
- TODESCO, J. L.; SOUZA, N. & MARTINS, L. C. (2002) - Implementação de um Data Warehouse para um Sistema de Avaliação Institucional - Estudo de Caso da Universidade do Vale do Itajaí. In: CONGRESSO BRASILEIRO DE COMPUTAÇÃO, Itajaí, SC. *Anais...* Itajaí, 2002.