UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Coleta de Requisitos e Modelagem de Dados para Data Warehouse: um Estudo de Caso utilizando Técnicas de Aquisição de Conhecimento

por

LUÍS CLÁUDIO CHAVES ZIULKOSKI

Projeto de Diplomação

Profª Drª Mara Abel
Orientadora

Porto Alegre, março de 2003.

Agradecimentos

Sumário

Lista de Abreviaturas

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CPD	Centro de Processamento de Dados
DW	Data Warehouse
ER	Entidade-Relacionamento
ProPesq	Pró-Reitoria de Pesquisa
SABi	Sistema de Automação de Bibliotecas
SAI	Secretaria de Avaliação Institucional
SQL	Structured Query Language
UFRGS	Universidade Federal do Rio Grande do Sul

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho é um estudo sobre o uso de técnicas de aquisição de conhecimento na coleta de requisitos e modelagem de dados para data warehouse.

São revisados conceitos básicos sobre data warehouse e enfatizadas as suas características como banco de dados de suporte à decisão. As técnicas utilizadas para a coleta de requisitos e modelagem de dados usadas tradicionalmente são apresentadas. A necessidade da coleta de requisitos é ressaltada e as características de modelos de dados para data warehouse são discutidas.

O processo de tomada de decisão baseado em data warehouse é enfocado no contexto da engenharia de conhecimento, procurando demonstrar como esta disciplina pode contribuir através do uso das técnicas de aquisição de conhecimento para a coleta de requisitos de data warehouses.

Para analisar a aplicabilidade das técnicas de aquisição de conhecimento na coleta de requisitos de data warehouse foi desenvolvido e avaliado um estudo de caso. O estudo de caso apresenta como o conhecimento adquirido se reflete no modelo de dados.

Palavras-chave: Armazém de Dados, Coleta de Requisitos, Modelagem Dimensional, Engenharia de Conhecimento, Aquisição de Conhecimento, Sistema de Apoio à Decisão

Abstract

1 Introdução

Saber fazer uso da informação é o grande diferencial competitivo das organizações atualmente. A Tecnologia da Informação é vista como um ferramenta estratégica central na busca pela vantagem competitiva [SHI02].

Com todo o processo de informatização, porém, a quantidade de dados disponíveis dentro das organizações chegou a um nível em que os tomadores de decisão podem acabar sufocados pela montanha de informações que recebem. Muitas vezes estas informações não são importantes ou são apresentadas de forma inesperada e incompreensível. O uso correto da informação depende do real entendimento sobre como se dá o processo de tomada de decisão numa organização.

Durante a década de 90 surgiu uma poderosa ferramenta para gerenciar as informações dentro de uma organização: o Data Warehouse, ou DW. As raízes da construção de DWs estão no aprimoramento das tecnologias de bancos de dados, que cada vez mais são capazes de manipular grandes volumes de dados. Muitas organizações construíram DWs como solução para integrar dados de diversos bancos de dados operacionais e suportar a tomada de decisão com informações qualificadas. Um DW é uma coleção de dados integrados, orientados por assunto, não-voláteis e variáveis com relação ao tempo, de apoio às tomadas de decisão gerenciais [INM96].

Em função das suas raízes como solução tecnológica, muitos DWs foram construídos internamente nos departamentos de sistemas de informação das organizações, partindo-se do pressuposto que as informações necessárias para a tomada de decisão já estão todas nos bancos de dados operacionais e que basta integrá-las e apresentá-las aos usuários para atingir um novo nível de suporte informatizado à decisão. Várias iniciativas de implantação de DW seguindo este paradigma resultaram em fracasso. Este é um erro já identificado na literatura [BAR97], mas que ainda ocorre em algumas iniciativas dentro das organizações.

Para implantar com sucesso um DW deve-se primeiro conhecer os requisitos dos futuros usuários, os tomadores de decisão e analistas da organização. Estes requisitos são de natureza diferente daqueles típicos dos sistemas que executam as atividades operacionais da organização. Estão ligados ao processo de tomada de decisão, que é uma função bem mais complexa do que a execução de procedimento diários e operacionais, e dependem muito mais do conhecimento e experiência da pessoa que a executa. Sendo assim, foi preciso desenvolver novas técnicas para coletar os requisitos para um DW.

Atualmente, a coleta de requisitos para DW vem sendo realizada utilizando técnicas como entrevistas e sessões de facilitação. Estas técnicas são muito dependentes de fatores humanos, como a capacidade de comunicação, e precisam de uma equipe com talentos especiais para terem sucesso. A coleta de requisitos em DW está mais para arte do que ciência [KIM98].

No domínio das tomadas de decisões organizacionais, a atividade dos usuários é intensiva em conhecimento e envolve aspectos como incertezas e *feeling*. Num domínio como este, as limitações da equipe de coleta de requisitos em fazer as perguntas certas e

as dificuldades do usuário em expressar todos os fatores envolvidos na sua tomada de decisão se somam e prejudicam o sucesso do projeto.

É preciso contornar estas dificuldades utilizando técnicas apropriadas para entender como se dá o processo de tomada de decisão e que informações são efetivamente utilizadas. Estas técnicas são encontradas numa disciplina emergente desta nova era que vivemos: a Engenharia de Conhecimento.

A Engenharia de Conhecimento lida com aquisição e representação de conhecimento e validação, inferência, explicação e manutenção de bases de conhecimento, segundo Turban, citado por [ABE02].

Aquisição de conhecimento refere-se à obtenção dos objetos, dos procedimentos e a forma geral de como o conhecimento é aplicado na solução de problemas em um domínio de aplicação. A tomada de decisão em uma organização é um exemplo de aplicação onde o conhecimento do especialista (o executivo ou gerente) é essencial na solução de problemas. Reconhecer as informações utilizadas pelo tomador de decisão corresponde a identificar os objetos do domínio e sua organização, num contexto de engenharia de conhecimento.

As técnicas de aquisição de conhecimento ajudam a reduzir os problemas de comunicação no processo de aquisição de conhecimento de fontes humanas [ABE01], como os tomadores de decisão, usuários de um DW.

Este trabalho é um experimento de aplicação de técnicas de aquisição de conhecimento na coleta de requisitos para um caso real de modelagem de DW. O objetivo do trabalho é estudar a aplicabilidade do uso de técnicas de aquisição do conhecimento para a coleta de requisitos e modelagem de DW.

No capítulo 2 são discutidos conceitos básicos de DW e há um aprofundamento sobre coleta de requisitos e modelagem de dados. No capítulo 3 são apresentados conceitos referentes à engenharia de conhecimento e descritas as técnicas de aquisição de conhecimento utilizadas neste trabalho. O capítulo 4 relata o desenvolvimento do estudo de caso e o capítulo 5 faz a avaliação do mesmo. Por fim, o capítulo 6 apresenta as conclusões e contribuições deste trabalho e dá sugestões de trabalhos futuros.

2 Coleta de Requisitos e Modelagem de Dados para Data Warehouse

O processo de tomada de decisão está continuamente evoluindo dentro de uma organização, à medida que novos desafios gerenciais são apresentados pela conjuntura econômica ou simplesmente pelo aprimoramento do processo nas organizações. Para se manter competitiva num mundo em que o *capital intelectual* é cada vez mais valorizado, uma organização necessita se manter bem informada. Mas se manter bem informada não consiste apenas em possuir montanhas de dados. Os dados precisam conter as informações certas na forma certa e na hora certa para acelerar e qualificar o processo decisório.

Para atender estas necessidades provenientes da evolução no processo de tomada de decisão e do aumento da competitividade no mercado, as organizações precisam dispor de um sistema automatizado de apoio à decisão. Um sistema que provê informações para seus usuários de forma que eles possam analisar uma situação e tomar decisões [POE98]. E que seja flexível para se adaptar às constantes mudanças impulsionadas pela evolução do processo de tomada de decisão.

Tipicamente, os usuários de sistemas deste tipo são gerentes, executivos e analistas de negócio. Eles desejam ter uma visão abrangente e distribuída ao longo do tempo de seus indicadores e métricas, que revelam o comportamento e o desempenho dos negócios. Um sistema de suporte à decisão deve, portanto, refletir os objetivos e metas da organização através da visão de seus tomadores de decisão.

A informatização nas organizações começou com a automatização das suas transações cotidianas, ou seja, do seu nível operacional. Ao longo dos anos, muitos sistemas transacionais foram desenvolvidos e junto com eles todo um universo de conhecimento, métodos, técnicas e tecnologias foi criado. Sempre houve uma esperança por parte dos profissionais de informática de que os mesmos bancos de dados construídos para suportar os sistemas transacionais pudessem ser utilizados para aplicações gerenciais. Mas essa esperança foi frustrada. Durante a década de 90 muitos estudos foram publicados demonstrando que o ambiente operacional, onde são executados os sistemas transacionais, é incompatível com o ambiente gerencial, onde se localizam os sistemas de apoio à decisão. Da mesma forma, os dados utilizados em cada um destes ambientes terão características distintas. A Tabela apresenta as principais diferenças entre os dados do ambiente operacional e os dados do ambiente gerencial:

Tabela 2.1 - Diferenças entre os dados de cada ambiente [INM96]

Dados do Ambiente Operacional	Dados do Ambiente Gerencial
Orientados por aplicação	Orientados por assunto
Detalhados	Resumidos ou refinados
Representam pontos no tempo (dados precisos no momento de acesso)	Representam valores ao longo do tempo (retrato dos resultados num período de tempo)
Sofrem atualização constante	Não são atualizados após inserção dos dados
Um registro acessado por vez	Conjunto de registros acessados por vez
Otimizados para transações	Otimizados para análise
Suportam operações cotidianas	Suportam necessidades gerenciais
Muitos usuários	Poucos usuários
Evita-se redundância (para otimizar transações)	Redundância é bem-vinda (para otimizar análises)
Estrutura estática, conteúdo variável	Estrutura flexível, conteúdo incremental

O ambiente operacional se refere basicamente à realização das transações cotidianas que fazem o negócio de uma organização funcionar. Já o ambiente gerencial se refere à avaliação do andamento do negócio. Como dito em [KIM98], o ambiente operacional é por onde as informações entram na organização e o ambiente gerencial é o onde as informações são consumidas (no processo de tomada de decisão).

Desde que a necessidade de uso de tecnologias diferentes das usadas até então no ambiente operacional se tornou fato reconhecido, tem sido feito esforço para elaborar e refinar tecnologias mais adequadas ao ambiente gerencial.

Estes esforços compreendem desde o desenvolvimento de novos mecanismos de processamento e armazenamento de dados até novas técnicas de coleta de requisitos e modelagem de informações.

Entre as novidades que surgiram para atender às demandas do ambiente gerencial, sem dúvida a tecnologia de DW é uma das que mais se destaca.

2.1 Revisão de conceitos sobre Data Warehouse

2.1.1 Definição de Data Warehouse

A definição mais difundida de DW é dada por [INM96]:

"uma coleção de dados integrados, orientados por assunto, não-voláteis e variáveis com relação ao tempo, de apoio às tomadas de decisão gerenciais."

O objetivo de um DW é fornecer informações que auxiliem no processo de tomada de decisão, descrevendo o comportamento da organização com dados históricos relacionados de forma significativa para a análise gerencial e estratégica.

Um DW deve [KIM98]:

- **Tornar as informações de uma organização acessíveis:** o conteúdo de um DW deve ser compreensível e navegável e o acesso aos dados deve ser feito com bom desempenho.

- **Tornar as informações de uma organização consistentes:** as informações oriundas de diferentes áreas da organização devem ter garantidas a integridade semântica, ou seja, combinadas sem problemas de nomes iguais para coisas diferentes, ou nomes diferentes para a mesma coisa.

- **Ser uma fonte de informações flexível e adaptável:** um DW deve suportar contínuas modificações estruturais para inserção de novos dados, que responderão a novas questões dos usuários, sem comprometer a estrutura já existente.

- **Ser o alicerce para a tomada de decisão:** o DW deve conter os dados certos na forma certa para suportar a tomada de decisão. *Só há uma verdadeira saída de um DW: as decisões que forem tomadas após o DW apresentar as evidências.* Um termo que antecede DW ainda é a melhor descrição do que um DW se propõe a fazer: um sistema de apoio a decisão.

Para atender a estes requisitos, um DW deverá ter as seguintes características [INM96]:

- **Integração:** Os dados que são inseridos no DW devem estar consistentes entre si em termos de nomes, formatos e unidades de medida. Em geral, os dados são provenientes de diversos sistemas de origem, internos ou até mesmo externos à organização, de tal forma que não existe padronização na nomeação de atributos, no formato de representação ou nas unidades de medida de valores numéricos. No processo de integração dos dados, também pode ser preciso corrigir dados que estejam inconsistentes na origem, devido à não-integração dos sistemas transacionais que provêem os dados.

- **Orientado por assunto:** um usuário de DW está interessado em temas importantes para o negócio da organização, então ele espera visualizar os dados por áreas (finanças ou vendas, por exemplo), encontrando os indicadores e métricas do negócio em cada área. No ambiente operacional, os bancos de dados são geralmente acessíveis apenas através de aplicações e, por isso, a estrutura de um banco de dados operacional reflete a divisão existente entre as aplicações. Já num DW, espera-se que o usuário trabalhe diretamente sobre o banco de dados e, por isso, sua estrutura deve refletir a divisão que os usuário estão acostumados a usar.

- **Não-volátil:** os dados que são inseridos no DW são estáticos. São dados que refletem situações consolidadas, que não sofrerão atualizações.

- **Variável com relação ao tempo:** os dados de um DW refletem situações ao longo do tempo, ou seja, são dados históricos por natureza. Esta característica é essencial do ponto de vista de avaliação do negócio da organização, pois é preciso manter séries históricas do comportamento dos negócios para que seja possível analisar os resultados mais recentes.

- **Suporte a grandes volumes de dados:** para realizar análises, avaliar o andamento do negócio e identificar tendências é preciso consultar um grande volume de dados. Um DW deve ser capaz de armazenar estes grandes volumes de dados e acessá-los com bom desempenho.

2.1.2 Arquitetura de Data Warehouse

Uma solução completa de DW deve contar com os seguintes três componentes [GON99]:

- Aquisição de Dados (*Back-End*): é o conjunto de programas utilizado para obter os dados que são inseridos no DW. Os programas providenciam conectividade com as fontes de dados, realizam limpeza, conversões de formato e correções de inconsistências para adequação dos dados ao modelo de dados do DW.

- Repositório de Dados: é o banco de dados propriamente dito. O armazenamento dos dados pode ser feito em bancos de dados relacionais, bancos de dados multidimensionais, ou uma solução híbrida, mantendo dados resumidos no banco de dados multidimensional e dados detalhados no banco de dados relacional. A estrutura dos dados no repositório corresponde ao modelo de dados do DW.

- Acesso aos Dados (*Front-End*): é o conjunto de produtos para acesso e análise de dados do DW. Permitem ao usuário navegar entre os dados, resumi-los, compará-los, etc. São ferramentas projetadas para trabalhar de acordo com o modelo de dados implementado. Portanto, a escolha do modelo de dados do DW é fator crítico para o front-end.

2.1.3 Abordagens para Desenvolvimento de Data Warehouse

A implantação de um DW em uma organização pode ser realizada sob duas principais abordagens, entre outras:

- Um DW único e global: o DW é projetado desde o início para conter todos os dados e servir a todos os setores da organização. É uma abordagem de alto custo e que demanda bastante tempo antes de apresentar resultados.

- Integração de Data Marts: um Data Mart é um DW de tamanho reduzido e de baixo custo, projetado para atender apenas a uma área da organização, mantendo informações sobre apenas um assunto (vendas, por exemplo). Construir um DW completo nesta abordagem seria um processo de criação de vários Data Marts de forma integrada. Para [KIM98] esta é a abordagem recomendada, desde que se obedeçam algumas restrições no projeto dos Data Marts para que a compatibilidade seja garantida num momento posterior.

2.2 Coleta de Requisitos para Data Warehouse

A coleta de requisitos para um DW não tem sido muito valorizada pelos autores da área. Historicamente, a construção do DW foi identificada como a etapa mais desafiadora na sua implantação. O processo de localizar a origem dos dados, integrá-los e organizá-los de maneira a tornar eficiente o acesso a um grande volume deles despertou maior interesse para os estudos. W. H. Inmon, que é considerado o criador do conceito de DW, inclusive propôs em [INM96] que o desenvolvimento de um DW

deveria seguir um "ciclo de vida invertido" com relação ao desenvolvimento de sistemas transacionais. Na sua visão, os requisitos para um DW *não podem ser conhecidos* antes de ele ser parcialmente carregado com dados e usado por analistas e tomadores de decisão. À medida que estes usuários utilizassem o DW eles informariam novas necessidades e quais dados seriam desnecessários, num processo de refinamento e crescimento da base de dados. Já no ciclo de vida de sistemas tradicional, supõe-se que os requisitos podem ser coletados e compreendidos antes do início do desenvolvimento, servindo então para orientar o desenvolvimento e estabelecer objetivos e metas, permitindo realizar planejamento e também controlar riscos no projeto.

Colocado desta maneira, desconsiderar os requisitos dos usuários ao iniciar um projeto parece arriscado. De fato, em [BAR97] encontramos um artigo em que a construção de um DW sem prévia coleta de requisitos é considerado um dos dez maiores erros em projetos de implantação de DW. Neste artigo, o autor relata que há vários casos de organizações em que a definição do conteúdo do DW foi orientada pela disponibilidade de dados nos sistemas que já operavam na instituição. Os responsáveis pela modelagem do DW oferecem os dados existentes aos usuários e questionam quais deles seriam interessantes para constar na base de dados. O resultado disto é um conjunto muito grande e complexo de informações, geralmente não organizadas conforme a visão dos usuários do DW e que acabam por se tornar incompreensíveis e inúteis. Esta impressão é compartilhada por [POE98] que afirma que a idéia de que simplesmente carregar um DW vai permitir que os usuários ganhem conhecimento do seu negócio é um dos mitos mais comuns na área, e que é também uma visão irrealista de um DW.

Autores posteriores a Inmon têm defendido a idéia da realização da coleta de requisitos logo no início de um projeto de implantação de DW em uma organização. [KIM98] e [POE98] descrevem ciclos de vida para DW que incluem uma etapa de coleta de requisitos como uma das primeiras a ser realizada e usada como base para as seguintes.

Mas a questão sobre como realizar a coleta de requisitos para o suporte a decisão em DW ainda está em aberto. Não há material em abundância sobre o tema e, infelizmente, toda experiência e conhecimento adquiridos no universo do ambiente operacional não é adequado para o ambiente de suporte a decisão.

Em [LAM95] o autor defende a idéia de quebrar com os velhos hábitos de eliciação de requisitos usados até então no ambiente operacional. Algumas características da eliciação de requisitos para DW segundo [LAM95] são:

- Entrevistar poucos usuários: no longo prazo, o DW deverá atender às necessidade de todos, mas no início é preciso começar com um conjunto reduzido de problemas a resolver, que trarão rápido benefício. Quanto mais usuários forem entrevistados, mais informações serão necessárias e diferentes formas de organização delas serão solicitadas, tornando o escopo do projeto muito grande, complexo e intratável. Ao contrário do ambiente operacional, onde várias pessoas executam as mesmas operações, no ambiente gerencial cada usuário tem sua particularidade.

- Evitar entrar em detalhes dos processos: a forma como os dados da organização são processados e o fluxo dos dados têm pouco a dizer sobre o

desempenho da organização. O foco da análise deve ser em cima dos indicadores da organização e dos dados necessários para calculá-los.

- Aceitar inclusão de novos requisitos a qualquer momento: num ambiente de suporte a decisão, será natural e indicativo de sucesso se os usuários começarem a pedir novas informações a partir do seu uso do DW. À medida que os usuários utilizam o DW eles percebem suas potencialidades e encontram novas possibilidades para seu uso. A equipe de desenvolvimento do DW deve, inclusive, incentivar mudanças nos requisitos.

Apesar da pouca quantidade de livros sobre coleta de requisitos em DW, existe material qualificado e detalhado sobre o tema em [KIM98], [POE98] e [KAC00]. O texto que se segue nesta seção é baseado nestes autores.

O principal propósito da coleta de requisitos para um DW é entender como os usuários conduzem seus negócios, que informações eles utilizam e o que eles gostariam de fazer no futuro.

Ao final do processo de coleta de requisitos deve-se obter:

- Um amplo entendimento dos negócios do usuário;
- detalhes específicos sobre os dados e os elementos principais para incluir numa implementação inicial do DW;
- um entendimento do uso essencial destes dados iniciais; e
- um entendimento das informações comuns que podem ser usadas por outras áreas da organização.

A forma preferida para realizar a coleta de requisitos é através de entrevistas orientadas por questões-chave, cujas respostas indicarão as informações utilizadas na organização para medir seu desempenho. Paralelamente pode ser feito um trabalho de levantamento da disponibilidade de dados junto ao departamento de sistemas de informação da organização, para que seja possível determinar a viabilidade e prazos realistas de implementação do modelo de DW a ser gerado, evitando criar expectativas não atingíveis nos usuários.

Outra forma de realizar a coleta de requisitos é utilizando sessões de facilitação. Numa sessão de facilitação um grande grupo de usuários (10 a 20, geralmente) é reunido num mesmo local e há uma pessoa, o facilitador, especialmente treinada para conduzir as discussões. Em geral, é preferível utilizar as sessões de facilitação num segundo momento da coleta de requisitos, para buscar atingir um consenso e uma padronização dos resultados obtidos num primeiro momento, através das entrevistas. Particularmente, quando um mesmo assunto diz respeito a várias áreas da empresa, o uso de sessões de facilitação se torna bastante atraente. Por fim, as sessões de facilitação são altamente recomendáveis para o planejamento da implantação do DW, definindo prioridades.

As razões da preferência por entrevistas são as seguintes:

- Facilidade de agendamento: como o número de usuários envolvidos em cada entrevista é pequeno, é mais fácil marcar os encontros do que se houvesse um grupo de dez ou mais pessoas.

- Estimular a participação: com poucas pessoas participando do encontro, cada participante se sente mais compelido a falar, garantindo-se assim que todas as vozes serão ouvidas.

Há dois componentes importantes para realização da coleta de requisitos através de entrevistas: a escolha dos entrevistados e a definição do conteúdo da entrevista.

2.2.1 Escolha dos Entrevistados

Para entendimento dos requisitos aos quais um DW deve atender, é preciso conversar com os futuros usuários do DW. Estes usuários são as pessoas dentro da organização que lidam com os rumos do negócio, são as pessoas que analisam a situação atual, observam o comportamento histórico dos negócios e decidem ou sugerem qual caminho a seguir para obter sucesso. Se localizam nas esferas gerenciais e executivas da organização, ou nos departamentos responsáveis pela avaliação dos negócios.

Mesmo que se esteja modelando um DW (ou Data Mart) para uma área de negócios específica da organização, é importante entrevistar algumas pessoas de outras áreas que tenham relação com a área alvo. Com isso, pode-se adotar uma estratégia de desenvolvimento integrada, mantendo a compatibilidade à medida que o DW vai incorporando novas área de negócio. É preciso obter um entendimento do vocabulário comum de toda a empresa para garantir que não se está modelando um DW que será uma ilha dentro da organização.

Podemos dividir os usuários que serão entrevistados nos seguintes grupos:

- Analistas da área alvo

- Gerentes da área alvo

- Analistas de áreas relacionadas

- Gerentes de áreas relacionadas

- Executivos

Deve-se entrevistar pessoas de diferentes níveis hierárquicos dentro da área alvo. Os executivos fornecem as estratégias de alto nível e a visão geral do assunto. Os gerentes de nível intermediário desvendam como as estratégias de alto nível são traduzidas em táticas do negócio e têm uma visão realista sobre onde gostariam chegar com mais informações e análise. Por fim, os analistas contribuem dizendo onde a informação está sendo usada dentro da organização.

Nas áreas relacionadas deve-se escolher alguns poucos representantes, preferencialmente aqueles que interagem com a área alvo. O foco nas conversas deve ser mantido nos dados que são compartilhados com a área alvo, para compreender as dependências existentes entre as áreas.

2.2.2 Conteúdo das Entrevistas [KIM98, POE98, KAC00]

2.2.2.1 Entrevistas com Executivos

O conteúdo das entrevistas com executivos é pouco extenso e mais geral, tratando principalmente de questões como missão e objetivo da organização, os obstáculos sobre os quais se deve passar para alcançar o sucesso e as medidas de sucesso. Entrevistas com executivos fornecem um entendimento macro da organização e de como ela é conduzida.

Sugestões de perguntas a fazer para o executivo são:

- Quais são os objetivos de sua organização? O que a organização procura alcançar?
- Quais são as métricas utilizadas para medir o sucesso da organização? Como saber se o negócio está indo bem? Com que freqüência as métricas são examinadas?
- O que pode impedir a organização de alcançar seus objetivos? Quais são os grandes desafios enfrentados?
- Como a organização identifica problemas? Como são tratados os problemas?

2.2.2.2 Entrevistas com Gerentes e Analistas

Estas entrevistas são similares às anteriores, porém num nível mais detalhado e restrito ao departamento ou setor onde o usuário atua. Aqui também são questionados objetivos, obstáculos, desafios e métricas, e deve ser feito um trabalho em cima dos relatórios manipulados pelos usuários, criados ou analisados. É com estes usuários que se obtém detalhes da descrição dos dados e de sua organização em níveis hierárquicos.

Sugestões de perguntas de alto nível para gerentes e analistas são as que seguem:

- Quais são os objetivos de seu departamento? O que o departamento procura alcançar?
- Quais são as métricas utilizadas para medir o sucesso do departamento? Como saber se o departamento está indo bem? Com que freqüência as métricas são examinadas?
- Quais são os grandes desafios enfrentados? O que limita o sucesso de seu departamento hoje?

Após as perguntas de alto nível deve-se partir para a descoberta dos requisitos de análise, questionando o usuário sobre os relatórios que ele recebe e utiliza, os relatórios que ele constrói, as análises *ad hoc* que ele eventualmente executa e quais são os seus desejos para melhorar seu trabalho. Abaixo está apresentada uma lista de perguntas sugeridas para cobrir estes tópicos:

- Que relatórios você recebe? Que dados no relatório são importantes? Como você usa esta informação?

- Que relatórios você cria? Que análises são executadas rotineiramente? De onde você obtém os dados? O que é feito com o relatório?

- Que tipos de análise *ad hoc* você costuma fazer? Quem solicita estas análises? Quanto tempo você leva para atender a uma demanda destas?

- Que análise você gostaria de fazer? Há aprimoramentos potenciais nos seus métodos atuais? Que oportunidades você visualiza para melhorar os negócios com um aprimoramento do acesso aos dados?

Outro tópico a cobrir com estes usuários é o entendimento dos detalhes descritivos dos dados, as hierarquias e agrupamentos existentes e o nível de detalhe em que são utilizados os dados. Sugestões de perguntas são as que seguem:

- Descreva seus produtos (ou outras dimensões importantes como clientes, fornecedores, ...). Como você agrupa os produtos em categorias?

- Em que nível de detalhe você costuma realizar suas análises? Em que período de tempo são feitas as análises? Quanto de informação histórica é necessário?

2.3 Modelagem de Dados para Data Warehouse

Na coleta de requisitos para o DW o objetivo era entender o negócio da organização e como ele é visto pelos membros da organização, identificando os dados utilizados no processo de tomada de decisão.

A modelagem de dados se preocupa em entender os dados, criando uma visão dos dados que nos permite compreender suas características e relacionamentos. A habilidade de visualizar algo tão abstrato quanto um conjunto de dados de forma concreta e tangível é o segredo da compreensibilidade [KIM96].

O modelo de dados utilizado em um DW é diretamente responsável por tornar as informações compreensíveis e navegáveis pelo usuário, um dos objetivos destacados na seção 2.1.1. O modelo de dados também será determinante para que o DW seja flexível e capaz de ser modificado para atender os incessantes novos requisitos que os usuários descobrem à medida que o utilizam. Por fim o modelo de dados utilizado interfere também no desempenho de acesso a grandes volumes de dados, como acontece no processamento analítico dos sistemas de apoio a decisão.

Um bom modelo de dados para DW terá as seguintes características:

- Reflete a visão que o usuário tem do negócio e suas medidas
- É simples
- É facilmente compreensível e memorizável para o usuário
- É flexível para incorporação de novos elementos
- Geram esquemas de banco de dados que permitem bom desempenho para consultas com grande volume de dados

Os modelos de dados utilizados no ambiente operacional não têm estas características. Pelo contrário, são muitas vezes o oposto. No ambiente operacional, o objetivo do modelo de dados é otimizar os dados para o processamento de transações. A redundância é o grande vilão para o processamento de transações, pois aumenta a complexidade de cada transação obrigando-a a atualizar dados em diversas estruturas e consumindo mais recursos. Além disso, quanto mais dados uma transação atualizar, mais tempo levará para ser concluída e mais estruturas de dados serão bloqueadas neste processo, acarretando sérios problemas para o ambiente concorrente de processamento em que usualmente os bancos de dados operacionais executam.

Em função desta necessidade de eliminar redundâncias nos bancos de dados operacionais, toda uma disciplina de modelagem de dados foi desenvolvida, utilizando a técnica de *normalização de arquivos* [HEU01]. Esta disciplina tem sido chamada de modelagem Entidade-Relacionamento (ER), apesar de um modelo ER não ter de ser normalizado por definição.

Um modelo ER normalizado é geralmente utilizado para implementar os esquemas de banco de dados, constituindo-se assim na visão que os usuários do banco de dados terão dos dados. Um modelo normalizado ilustra os relacionamentos microscópicos entre os dados, destacando as relações de dependência de dados e acabando por gerar uma grande quantidade de entidades e relacionamentos entre as entidades. Segundo [DEM01], a normalização tem muitas vantagens para manter a integridade dos dados de forma eficiente para o processamento on-line de transações, mas atua fortemente contra a legibilidade dos dados no ambiente de suporte à decisão, como segue:

- Gera um grande número de tabelas: o propósito da normalização é separar dados independentes em entidades distintas. É comum em atividades complexas, como fábricas, encontrar bancos de dados com mais de mil tabelas.
- Há múltiplas formas de ligar duas tabelas, e a escolha do caminho faz grande diferença no resultado obtido e no tempo de resposta de uma consulta.

Outra dificuldade com modelos ER é que eles são muito simétricos, ou seja, não é possível identificar as entidades importantes apenas baseando-se nas suas características no modelo, como os seus relacionamentos com outras entidades. Não há nenhuma propriedade relativa a forma de se relacionar com as demais entidades que

permita diferenciar entidades quanto à sua finalidade, que permita identificar construções regulares no diagrama.

Em resumo, um modelo ER normalizado é complexo e de difícil compreensão e memorização, características opostas às desejadas para um modelo de dados para DW. Além disso, como um modelo ER é irregular, as ferramentas de front-end não conseguem presumir nada a respeito do modelo de dados, sendo impedidas de utilizar interfaces padronizadas para o acesso aos dados.

Um esquema de banco de dados gerado a partir de um modelo ER normalizado dificilmente terá um desempenho adequado para responder consultas com grande volume de dados. As operações de ligação (ou junção) de tabelas em bancos de dados relacionais são custosas, e num modelo normalizado são necessárias muitas ligações entre tabelas. Para que o tempo de resposta de consultas como as usadas para análises seja prático, é preciso um esforço de sintonia fina da consulta pelo pessoal de administração do banco de dados. Em alguns é preciso até mesmo criar estruturas como índices secundários para possibilitar que a consulta apresente resultados num tempo de resposta aceitável. Num ambiente de DW, onde se espera que os usuários construam suas próprias consultas *ad hoc* ao banco de dados, a necessidade desta intervenção da administração de banco de dados é impraticável.

A conclusão a que se chega é que os modelos ER normalizados são muito bons para a atualização dos dados, mas ruins para consultá-los.

Com todas estas deficiências que a modelagem ER apresenta, foi preciso estabelecer um novo paradigma de modelagem de dados para DW. A proposta que vem se estabelecendo como preferencial é a modelagem dimensional (ou multidimensional), que foi lançada na década de 60 [RAD96] e permaneceu em segundo plano por longos anos, até que foi recuperada na década de 90 por Ralph Kimball [KIM96], seu principal defensor.

2.3.1 Modelagem Dimensional

Modelagem dimensional é uma técnica para construir modelos de negócio como conjuntos de medidas descritas através das diferentes facetas do negócio [RAD96]. É o nome de uma técnica de projeto que procura apresentar os dados numa estrutura padronizada que seja intuitiva e permita acesso com alto desempenho [KIM98].

A origem do termo *dimensional* está relacionada com a idéia de que os dados devem ser agrupados de maneira a formar um cubo, ou hipercubo, que seria a estrutura padrão para visualizar os dados.

Por exemplo [KIM96], se um executivo descreve as atividades de sua empresa da seguinte forma: “vendemos produtos em várias lojas e avaliamos nosso desempenho ao longo do tempo”, podemos imaginar esse negócio como um cubo de dados, como mostra a Figura . Os pontos dentro do cubo representam medidas do negócio válidas para a combinação de valores em cada uma das dimensões do cubo: foram vendidos 100 quilogramas de Arroz no Mercado Bom Fim durante o mês de fevereiro de 2003.

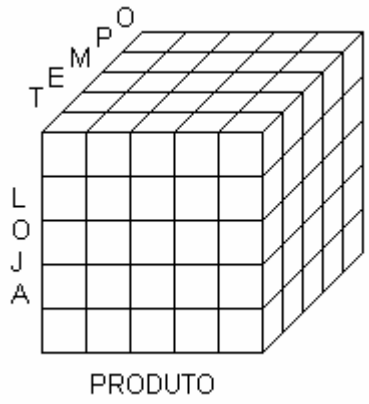


Figura 2.1 - O modelo dimensional de um negócio: cada ponto interno ao cubo contém as medições para uma combinação de Produto, Mercado e Tempo [KIM96]

Sobre o cubo são definidas operações como:

- *Slice and Dice*: ou fatiamento do cubo, é a restrição das coordenadas nas dimensões de acordo com critérios definido em cima de atributos das dimensões. Por exemplo: visualizar somente os dados dos mercados de pequeno porte.

- *Agregação*: permite reduzir a dimensionalidade de um cubo ou de uma fatia de cubo. Por exemplo: ao visualizar a receita total de cada mercado por mês estamos agregando os dados de receita na dimensão produto, eliminando-a do cubo.

- *Drill-up/Drill-down*: é a navegação entre níveis de agregação, de acordo com hierarquias existentes nas dimensões. Exemplo: o usuário pode começar visualizando dados totais de receita para os mercados e fazer um drill-up para visualizar os dados totais por bairro, agregando os mercados próximos.

O cubo é uma abstração poderosa para visualizar os dados, mas ele é restrito a três dimensões. Um modelo dimensional não pode, portanto, ser definido como sendo um cubo. É preciso estender esta visualização para *n* dimensões e a nossa mente não é capaz de construir imagens com mais de três dimensões. Por isso, um modelo dimensional é definido em termos de suas *dimensões*, com seus atributos e hierarquias, e em termos de suas medidas, ou *fatos*, que estão nos pontos de interseção das dimensões.

2.3.1.1 Fatos

Os fatos são medições do negócio. Geralmente são dados numéricos e aditivos, ou seja, podem ser agregados por soma, média ou outras funções. A aditividade dos fatos é uma propriedade muito desejável, pois permite que os dados sejam resumidos, o que é essencial quando se está lidando com grandes volumes de dados. Exemplos de fatos aditivos são: valor total da venda, quantidade de unidades vendidas.

Mas há também fatos não-aditivos e fatos semi-aditivos. Fatos não-aditivos são de pouca utilidade, só podem ser resumidos utilizando contagem. E fatos semi-aditivos são fatos que só podem ser agregados em algumas dimensões.

Uma observação importante a fazer é que não é obrigatório que todos os fatos sejam medidos em todas as interseções das dimensões. Isto é, não é preciso ter um valor para o fato para cada combinação de valores das dimensões. Na prática, não existem fatos para a imensa maioria das combinações, o que significa que o cubo é altamente esparso.

Um atributo do fato que é de fundamental importância é o *grão* do fato, isto é, o nível mais detalhado que se deseja para o fato. No exemplo do cubo, pode-se decidir por um grão correspondente ao total diário de vendas para cada produto em cada loja.

2.3.1.2 Dimensões

Uma dimensão é um conjunto de objetos que descrevem e classificam os fatos através de seus atributos. Os atributos de uma dimensão Loja, por exemplo, podem incluir uma hierarquia de bairro, cidade e estado e atributos descritivos como o nome do estabelecimento. Geralmente uma dimensão corresponde a um objeto, processo ou evento real para os quais existem dados.

São os atributos e as hierarquias das dimensões que permitem realizar operações interessantes sobre os fatos, como as operações de fatiamento, agregação e navegação hierárquica descritas anteriormente. Por isso, a descoberta dos atributos das dimensões é fator crítico para o sucesso de um DW.

Em geral, os atributos de dimensão são dados textuais. Quando um valor numérico é obtido nas fontes de dados pode existir dúvidas se ele é um atributo de dimensão ou um fato. Uma forma de resolver esta dúvida é observar se ele se modifica para cada combinação das dimensões. Se o valor se mantém constante ao longo de uma ou mais dimensões ele não é um fato, é um atributo de dimensão. Por exemplo, o tamanho de um produto é um valor numérico que não deve ser considerado um fato, pois dificilmente se modificará ao longo do tempo ou para diferentes mercados.

2.3.1.3 Esquema Estrela

A representação gráfica de um modelo dimensional não tem ainda uma notação própria amplamente adotada. Existem propostas de modelos dimensionais conceituais com notação própria, como a de Golfarelli [GOL98], chamada de *Dimensional Fact Model (DFM)*, onde todos os elementos do modelo têm notação própria (fato, dimensão, hierarquia, atributo, aditividade), mas em geral são usados modelos lógicos com notação semelhante a ER para representar modelos dimensionais.

Segundo [MOR98] são duas as razões principais para este desinteresse pela modelagem conceitual:

- O DW é voltado ao ambiente de negócios e seus usuários geralmente não dão importância a problemas conceituais.

- Modelos lógicos e físicos têm enfoque na implementação e otimização do desempenho do sistema, que é um dos objetivos do DW.

Uma das representações lógicas mais comumente utilizada é o *esquema estrela (star schema)*. Este esquema utiliza a abordagem relacional com algumas importantes restrições para representar o modelo dimensional. São usados os mesmos componentes de um diagrama ER lógico, como entidades, relacionamentos, chaves primária e estrangeira, cardinalidade etc.

No esquema estrela os fatos são representados na *Tabela de Fatos*, que é única num diagrama e ocupa a posição central. As dimensões são representadas cada uma em sua *Tabela de Dimensão*, que se posicionam ao redor da tabela de fatos, com a qual deve existir um relacionamento um-para-muitos. Esta disposição cria o padrão radial que dá nome ao esquema. Na Figura 2. é apresentado um exemplo de esquema estrela para o cubo apresentado anteriormente:

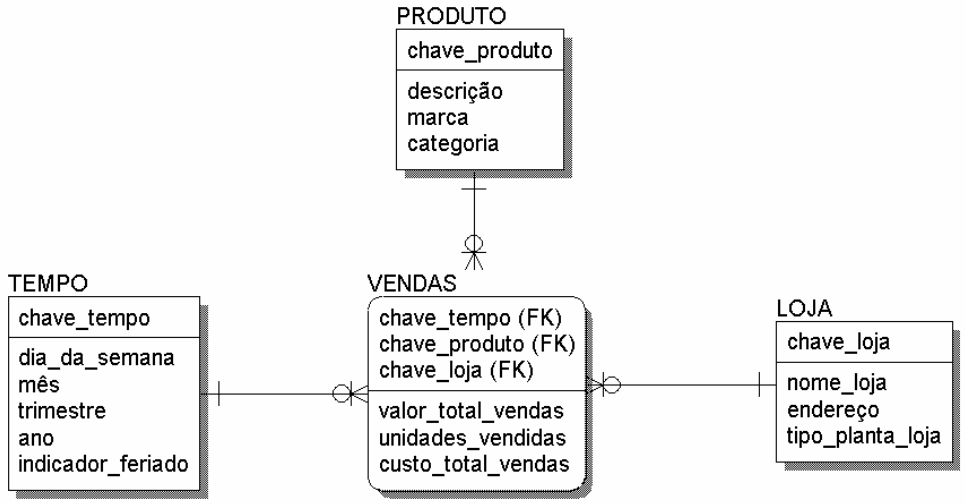


Figura 2.2 - Exemplo de esquema estrela

A chave primária da tabela de fatos é formada pelo conjunto das chaves estrangeiras provenientes das dimensões, de tal forma que ela é sempre a expressão de um relacionamento *n-ário* entre todas as dimensões. Os únicos relacionamentos permitidos são entre tabela de dimensão e tabela de fatos, ou seja, não pode haver ligação direta entre dimensões.

Os fatos são os atributos da tabela de fatos. Não há representação explícita para as hierarquias no esquema estrela. Elas são implicitamente representadas pelos atributos das tabelas de dimensão. A aditividade também não é representada. As hierarquias e aditividades devem ser especificadas com documentação extra.

Existe uma variação do esquema estrela conhecida com *esquema floco-de-neve (snowflake schema)*, que faz uma normalização nas tabelas de dimensão. O esquema

floco-de-neve evita a redundância de atributos descritivos nas dimensões, criando tabelas auxiliares para armazenar estes atributos cujos valores costumam se repetir muito na dimensão. Num esquema floco-de-neve a hierarquia de uma dimensão seria explicitada pela criação de tabelas auxiliares para cada nível superior.

Em geral, normalizar uma dimensão traz pouco ou nenhum benefício para o modelo [KIM98]. Em alguns casos pode até piorar a compreensibilidade para o usuário. Depende de uma avaliação criteriosa a decisão sobre normalizar ou não os atributos em uma dimensão, caso o projetista do DW considere que esta opção é vantajosa por que vai trazer melhoras no desempenho de consultas e atualizações e no entendimento do usuário não há problemas em normalizar uma dimensão. Mas, como regra geral, normalizar dimensões para usar um esquema floco-de-neve é dispensável.

Um modelo dimensional implementado no esquema estrela possui todas as características desejáveis para um modelo de dados para DW: é bastante simples, sua analogia com um hipercubo de dados facilita a compreensão pelos usuários e pela sua simplicidade é possível memorizar os elementos mais importantes. É um modelo que facilmente suporta inclusão de novos elementos, como novas dimensões, novos atributos ou novos fatos, sem comprometer a estrutura existente. Finalmente, sua implementação tanto em bancos de dados multidimensionais ou bancos de dados relacionais tem um desempenho muito bom.

Há também outras vantagens no modelo dimensional com esquema estrela [KIM98]. Primeiro, a estrutura padronizada e regular de um modelo como esse permite que tanto o sistema de gerência do banco de dados quanto as ferramentas de acesso aos dados assumam premissas sobre os dados que facilitam a apresentação e o desempenho. Segundo, há um bom número de abordagens padrão para resolver problemas comuns no mundo dos negócios. E terceiro, a grande maioria dos softwares e pacotes para navegação e agregação de dados dependem de uma estrutura de fatos e dimensões do modelo dimensional.

Mas há também problemas com o esquema estrela. O principal é o limitado poder de expressão, devido às suas restrições. Algumas pessoas inclusive alegam que o esquema estrela só funciona em negócios de varejo. Embora esta afirmativa seja falsa, é verdade que em alguns negócios é preciso relaxar algumas restrições do esquema estrela e fazer modelos dimensionais estendidos, como o exemplo apresentado na seção 2.3.1.6, onde é necessário manter um relacionamento muitos-para-muitos entre uma dimensão e a tabela de fatos.

Com relação ao desempenho quando implementado sobre bancos de dados relacionais, o esquema estrela pode ficar prejudicado em casos que uma ou mais dimensões tenham cardinalidade muito grande (muitos registros), porque isto tornará as junções mais custosas. Este seria um caso para considerar a normalização da dimensão. No esquema estrela também não há estruturas previstas para armazenar dados já agregados e resumidos, uma das principais estratégias para otimização do acesso aos dados.

2.3.1.4 Projeto de um Modelo Dimensional

Em [KIM98, KIM96] há uma proposta de método para projeto de um modelo dimensional seguindo quatro passos, que são escolhas consecutivas:

1. Escolha do processo do negócio que será coberto pelo modelo, ou seja, a identidade da tabela de fatos.

2. Escolha do nível de detalhe (grão) da tabela de fatos.

3. Escolha das dimensões.

4. Escolha dos fatos.

Presume-se que um trabalho de coleta de requisitos como o descrito anteriormente já tenha sido realizado, identificando quais são os assuntos, ou processos do negócio da organização, que serão inseridos no DW e que informações são necessárias para cada um.

Um modelo de dados completo para DW é composto por vários modelos dimensionais, cada um para uma tabela de fatos. Em um DW realmente integrado, espera-se que as dimensões que se repetem nos diversos modelos estejam em conformidade, ou seja, tenham a mesma definição. Este é o fundamento da arquitetura de DW proposta em [KIM98], onde estes vários modelos correspondem a um Data Mart cada, e geralmente seus fatos se originam de um único sistema transacional.

O projeto de um modelo dimensional se inicia pela escolha de um processo do negócio. Exemplos de processos para uma empresa do ramo de varejo são: pedidos de compra, pedidos de cliente e expedição para cliente.

Após a identificação do processo deve ser definido precisamente o que representa cada registro na tabela de fatos, ou seja qual é o nível de detalhe, o grão da tabela de fatos. Níveis de detalhe típicos incluem uma transação de venda individual, um item de uma ordem de compra, um instantâneo de totais diários ou mensais. Recomenda-se que seja utilizado o menor grão possível, pois o projeto será mais robusto, suportando novas consultas inesperadas e adição de novos elementos do modelo mais facilmente.

Conhecendo a granularidade da tabela de fatos a definição das dimensões é quase direta. O próprio grão determinará um conjunto primário ou mínimo de dimensões, sem as quais não é possível caracterizar o grão. Por exemplo, para um item de uma fatura as dimensões de data, cliente, produto e a dimensão degenerada[1] número da fatura são imediatas. O projetista pode então adicionar outras dimensões descritivas que sejam úteis, examinando as informações descritivas disponíveis sobre o grão da tabela de fatos, dirigido pelos requisitos do negócio. A escolha das dimensões é o passo-chave do projeto. Deve-se decidir sobre a identidade das dimensões antes de se preocupar com o local onde serão obtidas as informações.

[1] Uma dimensão degenerada é uma dimensão que só tem a sua chave como atributo, dispensando a necessidade de uma tabela para representá-la. Uma dimensão degenerada é representada apenas pela presença de sua chave na chave primária da tabela de fatos.

Por fim, são escolhidos os fatos que serão colocados na tabela de fatos. Para uma granularidade correspondente a uma transação geralmente só há um fato para medir, que é o total da transação. Já com a granularidade de um instantâneo de totais num período, é provável que haja muitos fatos pois qualquer resumo de atividade pode ser um fato útil. Os fatos devem ser sempre relativos ao grão da tabela. Não se deve misturar fatos relativos a outros períodos de tempo ou agregações para facilitar os cálculos. As agregações são bem-vindas, mas devem, ser armazenadas em tabelas de fato diferentes, com o grão correspondente.

2.3.1.5 Tabelas de Fatos sem Fatos [KIM96, KIM98]

Em algumas situações pode acontecer de não ser encontrado nenhum fato mensurável para ser armazenado na tabela de fatos. Isso é normal e já foi identificado em vários processos de negócio merecedores de um DW. As tabelas de fatos para estes modelos são chamadas *tabelas de fatos sem fatos*. Dois casos são os mais comuns: tabelas de rastreamento de eventos e tabelas de cobertura. Em tabelas de fatos sem fatos o fato é a ocorrência ou não do processo que o grão da tabela representa. A forma de resumir dados para tabelas de fatos sem fatos é através de contagens.

Um exemplo onde se usa rastreamento de eventos é para registrar a freqüência em faculdades. O grão da tabela de fatos é a presença do aluno em sala de aula. Dimensões possíveis seriam data, estudante, curso, professor e local. Mas não há fato. Não há medidas que possam ser feitas sobre o evento da presença do aluno. Apesar disso, consultas interessantes podem ser realizadas, como:

- Que aulas têm maior freqüência?
- Que professores dão aulas para as maiores quantidades de alunos?
- Que locais são mais usados?

2.3.1.6 Dimensões com relacionamento muitos-para-muitos [KIM98]

Num modelo dimensional, a relação esperada entre as dimensões e a tabela de fatos é sempre de um-para-muitos, ou seja, um objeto da dimensão terá muitos registros relacionados na tabela de fatos, variando ao longo das demais dimensões. Mas há casos em que um registro da tabela de fatos está indiscutivelmente ligado a um número variável de objetos de uma dimensão, que não pode ser determinado *a priori*.

Um exemplo é na área bancária. Suponha que há um tabela de fatos para armazenar o balanço mensal das contas-correntes. Cada conta-corrente pode ter mais de um cliente do banco como correntista. Um cliente do banco pode ter mais de uma conta também. Nitidamente, o relacionamento entre cliente e conta-corrente é muitos-para-muitos.

Na abordagem relacional, relacionamentos muitos-para-muitos são resolvidos através da criação de uma tabela intermediária, com relacionamento muitos-para-um com as duas tabelas originalmente relacionadas, e com a chave primária da tabela intermediária sendo formado pelas chaves estrangeiras das tabelas relacionadas.

Esta solução pode ser adotada na implementação do modelo dimensional sobre um banco de dados relacional, desde que sejam tomadas precauções para evitar que a junção das tabelas não gera resultados incorretos.

Uma conseqüência direta de um relacionamento muitos-para-muitos entre uma tabela de dimensão e a tabela de fatos é que sempre que for realizada a ligação destas tabelas um registros da tabela de fatos será repetido tantas vezes quanto o número de objetos na dimensão relacionados. Dessa forma, registros de balanço de conta-corrente poderiam ser usados mais de uma vez na hora de realizar um agregação, gerando resultados incorretos.

Um artifício para contornar este problema é o uso de pesos na tabela intermediária, indicando a contribuição percentual de cada objeto na dimensão para cada fato na tabela de fatos. Este peso multiplica o valor do fato quando é feita a ligação das tabelas, garantindo que o total agregado esteja correto.

Na Figura está apresentada a resolução do relacionamento muitos-para-muitos do exemplo bancário.

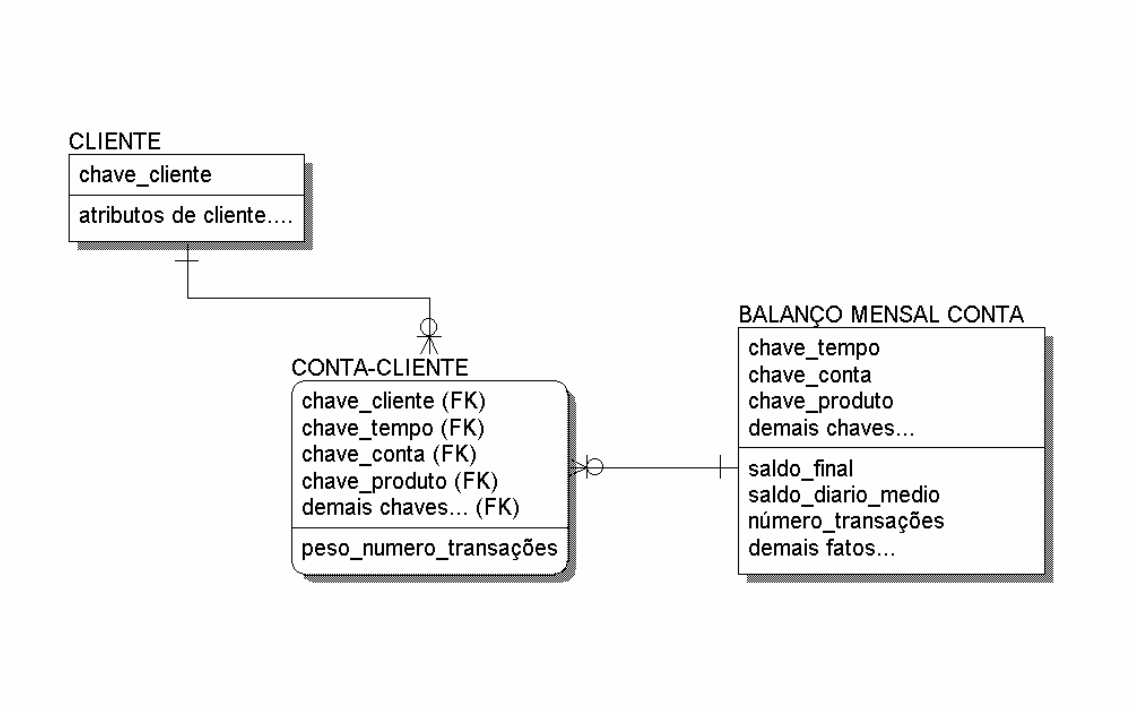


Figura 2.3 - Resolução de relacionamento muitos-para-muitos no modelo dimensional

3 Aquisição de Conhecimento

Técnicas de aquisição de conhecimento são parte de uma disciplina chamada Engenharia de Conhecimento [SCH99]. A engenharia de conhecimento agrega um conjunto de metodologias, técnicas e formalismos que suportam a construção de sistemas de conhecimento. O objetivo geral da engenharia de conhecimento assemelha-se ao da Engenharia de Software: transformar o processo *ad hoc* de construir sistemas baseados em conhecimento em uma disciplina de Engenharia.

A engenharia de conhecimento lida com aquisição e representação de conhecimento e validação, inferência, explicação e manutenção de bases de conhecimento.

- **Aquisição de conhecimento** refere-se a obtenção de conhecimento de suas diversas fontes: livros, documentos, computadores com principal ênfase, por sua dificuldade, a especialistas humanos. O conhecimento refere-se a solução de problemas em um domínio de aplicação e envolve os objetos do domínio, os procedimentos e a forma geral de como o conhecimento é aplicado.

- **Representação de conhecimento** refere-se a escolha de uma forma de representação que possa compor um modelo do domínio e a codificação da informação adquirida nessa forma. Refere-se habitualmente ao conhecimento declarativo.

- **Validação do Conhecimento:** Verificação da consistência da base de conhecimento.

- **Inferência**: Definição dos procedimentos de manipulação e aplicação do conhecimento, com fins de implementação.

- **Explicação e Justificativa**: Envolve a recuperação do raciocínio do sistema ao atingir determinada conclusão e a definição da forma de apresentar esses caminhos de inferência para o usuário.

3.1 Conceitos relativos a Engenharia de Conhecimento

Antes de apresentar as técnicas de aquisição de conhecimento é preciso esclarecer alguns conceitos relativos a engenharia de conhecimento.

3.1.1 Conhecimento

O primeiro conceito a definir é o de conhecimento, e a diferença entre conhecimento, informação e dado [ABE02]:

- **Conhecimento:** consiste em (1) descrições simbólicas que caracterizam os relacionamentos empíricos e definicionais em um domínio e (2) os procedimentos para manipulação dessas descrições. Conhecimento inclui a

informação sobre o domínio e a forma como essa informação é utilizada para resolver problemas. Ex.: Maria tem mais de 18 anos. Maiores de 18 anos são responsáveis legais por seus atos. Maria será cobrada pelos danos.

- **Informação:** Reconhecimento dos objetos do domínio, suas características, suas restrições e seus relacionamentos com os outros objetos, sem ater-se a *utilidade* dessa informação. É o dado com o seu significado associado. Ex.: Idade de Maria = 20 anos

- **Dado:** Representação simbólica de um objeto ou informação do domínio sem considerações de contexto, significado ou aplicação. Ex.: 20 anos

- **Domínio:** Qualquer conjunto relativamente circunscrito de atividades.

3.1.2 Sistemas de Conhecimento

Um sistema de conhecimento é qualquer sistema de informação que gerencie, armazene e/ou aplique conhecimento organizacional explicitamente representado. O conhecimento pode ser de fonte humana, de outros sistemas, de livros, etc. O termo inclui sistemas especialistas, sistemas baseados em conhecimento, bancos de dados inteligentes e sistemas de informação intensivos em conhecimento, que possuem em comum o fato de modelarem conhecimento de forma explícita (não embutido ou disperso nos algoritmos do sistema) e aplicá-lo no suporte a solução de problemas.

Os benefícios do ponto de vista da organização que utiliza sistemas de conhecimento acontecem em diferentes áreas:

1. Produtividade: tomada de decisão mais rápida, aumento da produtividade, melhora no processo de solução de problemas, solução de problemas complexos, confiabilidade das decisões, automatização na operação de equipamentos.

2. Preservação do conhecimento organizacional: captura da perícia, organização do conhecimento disperso, reuso do conhecimento.

3. Disseminação do conhecimento organizacional: possibilidade de utilização do conhecimento longe de suas fontes, padronização das soluções aplicadas em qualquer ponto da organização.

4. Qualidade da decisão: aumento de qualidade das decisões, possibilidade de tratar com incertezas.

5. Treinamentos: melhora da qualidade dos treinamentos de funcionários e clientes pelo uso do conhecimento preservado (casos, acesso a informação, entre outros).

É interessante para este trabalho observar que o uso de um DW numa organização gera alguns dos mesmos benefícios, o que demonstra que existe uma certa afinidade entre DW e sistemas de conhecimento. Um DW também visa a uma tomada de decisão mais rápida, confiável e qualificada. E utilizando ferramentas de acesso a dados que permitem salvar as consultas utilizadas e compartilhá-las com outros usuários

é uma forma indireta de preservar e disseminar conhecimento, pois evidencia as informações utilizadas no processo de tomada de decisão e estimula os usuários a também compartilhar suas experiências e critérios. Porém um DW não manipula conhecimento explicitamente representado e por esse motivo não pode ser enquadrado como sistema de conhecimento.

3.2 Técnicas de Aquisição de Conhecimento

Originalmente utilizadas na construção de sistemas de conhecimento, cada vez mais as técnicas são utilizadas para dirigir entrevistas com usuários e clientes para coleta de requisitos na construção de sistemas convencionais, obter informações dos trabalhadores de conhecimento para modelar organizações, ou simplesmente transferir conhecimento entre diferentes profissionais [ABE01].

São especialmente bem-vindas quando o conhecimento envolvido na solução de problemas é mais tácito do que explícito, pois *"podemos saber mais do que podemos dizer"* segundo Polanyi, citado por [ABE01], o que se reflete na capacidade de explicitar e transmitir esse conhecimento. O conhecimento tácito é pessoal, específico ao contexto e difícil de ser formulado e comunicado. Já o conhecimento explícito ou "codificado" refere-se ao conhecimento transmissível em linguagem formal e sistemática.

Neste trabalho foram utilizadas duas classes de técnicas: entrevistas e classificação de termos, que serão detalhadas mais adiante. Porém há uma variedade de outras técnicas que merecem ser citadas:

- **Análise de Protocolos:** o especialista deve verbalizar seus pensamentos durante a execução de um processo, decompondo-o em subprocessos. É uma técnica bastante eficaz para especificação de tarefas e inferência.

- **Focalizando contextos ou cenários:** consiste na apresentação ao especialista de casos de teste reais ou criados pela engenheiro, procurando evidenciar pontos essenciais do método de solução do problema, como necessidades de tratamento de incertezas, por exemplo.

- **Observação:** corresponde a assistir ao especialista enquanto ele executa uma tarefa, real ou simulada. Pode incluir interrupção do engenheiro para fazer perguntas, mas geralmente tem melhores resultados sem interrupções. Garante uma visão realista do processo de solução no ambiente onde ele ocorre, mas não permite compreender porque as decisões foram tomadas, sendo necessário esclarecê-las com entrevistas posteriores.

- **Recuperação de Eventos:** utilizada para recuperar casos não triviais, ou de exceção, na solução de problemas. O especialista deve descrever o caso e a solução proposta.

3.2.1 Entrevistas [SCH99]

Praticamente toda a eliciação de conhecimento começa por entrevistas. Esta é a técnica mais comumente utilizada para aquisição de conhecimento e tem as mais

variadas formas, desde entrevistas completamente livres ou não-estruturadas até formalmente planejadas e rigidamente estruturadas.

Entrevistas não-estruturadas não têm pauta definida nem pelo especialista nem pelo engenheiro de conhecimento (ou, pelo menos, não tem pauta detalhada). Naturalmente, isto não significa que o engenheiro não tem objetivos definidos para a entrevista, significa apenas que não há restrições e que o escopo da entrevista é livre.

As vantagens desta abordagem provêem desta liberdade. Primeiro, ela pode ser usada sempre que se desejar estabelecer um bom relacionamento entre engenheiro e especialista. Segundo, o engenheiro pode facilmente ter uma visão ampla do tópico da entrevista. E terceiro, o especialista pode descrever o domínio da forma como lhe é familiar, enfatizando os tópicos que considera mais importante e ignorando os que considera desinteressantes.

As desvantagens também são devidas à falta de restrições. A ausência de uma pauta pode levar à ineficiência, pois o especialista pode ser desnecessariamente prolixo ou se concentrar apenas em tópicos cuja importância ele exagera. O especialista também pode ser muito tímido e economizar nas palavras. O domínio é descrito de uma forma muito desconexa, em diversos retalhos. Pode ser muito difícil integrar os dados, seja porque eles não forma uma unidade, ou porque há inconsistências. Esta última desvantagem ocorrerá com mais freqüência se as informações forem obtidas de diversos especialistas.

Na abordagem com entrevistas estruturadas o engenheiro de conhecimento planeja os tópicos a serem discutidos e conduz a conversa de acordo com o seu planejamento. Existe uma pauta devidamente estruturada, que implicará em uma transcrição também estruturada da entrevista, facilitando a análise dos dados.

Um problema com as entrevistas, estruturadas ou não, é que o especialista só vai transmitir o conhecimento que ele consegue verbalizar. Se houver aspectos não-verbalizáveis no domínio a entrevista não será capaz de descobri-los. Um aspecto pode ser não-verbalizável por duas causas. Primeiro, porque o conhecimento pode nunca ter sido representado ou articulado em termos de linguagem (é o caso de reconhecimento de padrões visuais, por exemplo). Segundo, o conhecimento pode até ter sido internalizado explicitamente de uma forma proposicional ou por algum tipo de linguagem, mas ao longo do tempo e com a experiência as decisões acabaram sendo mecanizadas. Isto pode ocorrer de uma forma tão intensa num especialista que ele pode passar a atribuir suas decisões mais complexas à sua intuição, apesar de, na verdade, estarem baseadas em uma grande quantidade de dados e experiências passadas que estão em sua memória, sempre relembradas pela contínua aplicação de estratégias. Numa situação destas é comum o especialista responder às perguntas com um "Eu não sei como faço isso..." ou "Era a coisa mais óbvia a ser feita...".

Outro problema é que as pessoas geralmente tentam justificar suas decisões da maneira que for possível. É comum o engenheiro de conhecimento observar que o especialista tomou uma decisão bastante válida, mas justificou de uma forma espúria.

Por essas e outras razões é que deve-se suplementar as entrevistas com métodos adicionais de eliciação. A eliciação deve sempre consistir de um programa de técnicas e métodos complementares.

Entrevistas não-estruturadas são geralmente utilizadas apenas nas etapas inicias da aquisição de conhecimento. As entrevistas estruturadas são particularmente úteis para refinamentos e para esclarecer dúvidas.

3.2.2 Classificação de Termos [SCH99]

Esta técnica é útil quando se deseja descobrir as diferentes visões que um especialista tem dos relacionamentos entre um conjunto de termos (ou conceitos). Basicamente é apresentado um conjunto de fichas contendo um conceito cada. As fichas são misturadas e solicita-se ao especialista que ele as separe segundo um critério qualquer a sua escolha. Pode-se solicitar que ele separe as fichas tanto num número fixo de subconjuntos quanto num número qualquer conforme ele considerar apropriado. Este procedimento é então repetido várias vezes, com o especialista utilizando critérios diferentes para classificar os conceitos, gerando subconjuntos distintos.

A cada classificação obtida deve-se questionar o especialista sobre o critério utilizado e sobre os nomes ou rótulos para os subconjuntos. É possível que o especialista resolva descartar conceitos porque são sinônimos ou acrescente conceitos que lhe parecem estar faltando.

Há muitas variações para esta técnica, dependendo dos conjuntos ou subconjuntos utilizados a cada nova classificação. Por exemplo, pode-se solicitar que o especialista subdivida um conjunto que resultou de uma divisão anterior, realizando uma classificação hierárquica. Ou então utilizar o conjunto completo de conceitos a cada nova classificação. Definir ou não o número de subconjuntos a ser gerado a cada classificação também são variações possíveis.

Os resultados obtidos com a classificação de termos são:

- Reconhecimento do relacionamento entre os termos, como hierarquias, atributos comuns e sinonímia.
- Obtenção de termos não evidenciados anteriormente.
- Reconhecimento de diferentes formas de visualizar o domínio.
- Uma melhor compreensão global do domínio.

As vantagens desta técnica são a rapidez para aplicar e a facilidade de análise dos resultados. Ela força que as construções mentais do especialista sejam colocadas num formato explícito. Muitas vezes ela é instrutiva até mesmo para o especialista, pois pode levá-lo a ver uma estrutura que ele mesmo utiliza mas que nunca havia articulado conscientemente antes.

Há alguns problemas potenciais com esta técnica. Um especialista pode acabar confundindo conceitos se não utilizar as mesmas distinções semânticas durante a sessão de trabalho. Também pode ocorrer do especialista simplificar demasiadamente a categorização dos termos, perdendo detalhes importantes.

Os termos utilizados para a técnica precisam ser determinados previamente, o que é usualmente realizado através da seleção de termos a partir de transcrições de entrevistas.

4 Estudo de Caso

A fim de experimentar o uso de técnicas de eliciação de conhecimento na coleta de requisitos e modelagem de DW foi desenvolvido um estudo de caso.

O estudo de caso foi baseado na avaliação da produção intelectual na Universidade Federal do Rio Grande do Sul (UFRGS). A produção intelectual é um dos principais indicadores de desempenho em uma universidade. Artigos científicos, capítulos de livros, trabalhos em anais, partituras musicais, apresentações artísticas, desenvolvimento de aplicativos computacionais e uma infinidade de outras formas de expressão do conhecimento e cultura são considerados produção intelectual.

A quantidade destes itens produzidos pela universidade, através do seu corpo docente e do seu corpo técnico fornece a medida da disseminação da cultura e do saber que a universidade está proporcionando à sociedade. Além disso, também é utilizada para medir o desempenho dos departamentos e docentes da universidade, possibilitando estabelecer índices para distribuição orçamentária, vagas em concurso público para novos docentes, bolsas de pesquisa, financiamento de projetos, etc.

Na UFRGS as informações sobre produção intelectual são encontradas em sistemas distintos e incompatíveis: o sistema de bibliotecas (SABi) e o currículo Lattes, fornecido pelo CNPq (Conselho Nacional de Pesquisa – órgão federal que coordena pesquisa no país). Além disso, os dados sobre os departamentos, os docentes e os técnicos, com os quais se deseja relacionar a produção intelectual se encontram em um outros sistemas, desenvolvidos internamente na instituição.

A simples falta de um repositório de dados que torne as informações sobre produção intelectual disponíveis de forma integrada já justifica a criação de um DW. Mas por parte da alta administração da universidade também há o interesse em solidificar um processo de avaliação institucional, sendo que inclusive existe uma secretaria especialmente criada para tal fim (Secretaria de Avaliação Institucional - SAI). Há também a intenção do setor de sistemas de informação, o Centro de Processamento de Dados (CPD), em empreender um projeto para criação de um DW para toda a universidade, justamente para dar suporte aos processos de decisão e à avaliação institucional. A criação de um DW seria o segundo passo em todo um processo de democratização da informação na universidade. O primeiro passo foi dar suporte informatizado para o ambiente operacional da universidade, fornecendo sistemas para controlar os diversos processos cotidianos da instituição, sendo que praticamente todos estes sistemas são integrados utilizando um mesmo banco de dados. Uma das poucas exceções é justamente o sistema de bibliotecas, principal registro da produção intelectual universitária.

Este trabalho está inserido no esforço do CPD para implantação de um DW completo para a universidade. Os resultados obtidos aqui poderão ser utilizados pela equipe do CPD para avaliar o uso de técnicas de eliciação do conhecimento no processo de modelagem do DW da universidade.

A escolha do assunto para este estudo de caso foi definida em conjunto com a equipe do CPD e a SAI, procurando ao mesmo tempo servir de base para

estabelecimento de uma metodologia de modelagem de DW na UFRGS e atender a uma demanda específica da SAI, que está desenvolvendo um levantamento de vários dados institucionais para o processo de avaliação da universidade.

Com o propósito de tornar o estudo de caso executável num período curto de tempo foi delimitado então um escopo bastante restrito, abordando-se apenas a questão da produção intelectual.

O desenvolvimento do estudo de caso constituiu-se das seguintes etapas:

1. Definição do método a aplicar
2. Escolha dos participantes na modelagem
3. Realização de entrevistas livres
4. Aplicação da técnica de Classificação de Termos
5. Elaboração do modelo de dados
6. Levantamento de origens dos dados presentes no modelo de dados

Nas próximas seções, cada uma das etapas será detalhada.

4.1 Definição do método a aplicar

Como na maioria dos projetos onde se aplica Engenharia do Conhecimento, o uso de entrevistas como técnica inicial foi adotado. O principal objetivo das entrevistas foi conhecer o domínio estudado (avaliação da produção intelectual da universidade), as informações necessárias para a tomada de decisão e coletar os termos mais importantes no domínio.

Na segunda fase, o relacionamento entre os termos destacados nas entrevistas foi investigado com o uso da técnica de classificação de termos. Os termos destacados representam os objetos do domínio a respeito dos quais se deseja manter informações. O entendimento das relações existentes entre estes objetos permite então construir um modelo de dados conforme a visão do especialista, que será o usuário do DW.

4.2 Escolha dos participantes na modelagem

A escolha das pessoas a entrevistar foi definida conjuntamente com a equipe do CPD. Uma vez que existia uma demanda por parte da SAI, foi natural que representantes desta secretaria fossem incluídos no projeto. Estes usuários poderiam ser enquadrados como sendo os analistas de negócios da organização, conforme definido na seção 2.2.1. Participaram da modelagem a diretora da secretaria e sua assistente.

Além da SAI, foi constatado que a Pró-Reitoria de Pesquisa (ProPesq) seria outro órgão de grande interesse na avaliação da produção intelectual, uma vez que este é um dos principais indicadores de desempenho da pesquisa e é um dado envolvido nos

processos de tomada de decisão na pró-reitoria. Para participar das entrevistas foi convidada então a vice pró-reitora de pesquisa.

4.3 Realização de entrevistas livres

Foram realizadas duas entrevistas livres, ou não-estruturadas, uma com a SAI e outra com a ProPesq, com duração de cerca de uma hora cada. O entrevistador procurou interferir minimamente na fala dos entrevistados, procurando sempre incentivar os entrevistados a falar sobre suas atividades e necessidades de informação, tanto presentes como futuras, desconsiderando quaisquer limitações quanto à indisponibilidade de dados.

A entrevista livre não seguiu um roteiro pré-definido. Havia apenas uma pergunta tema principal:

> *"Numa situação hipotética e ideal, onde todos os dados imagináveis estariam disponíveis, que informações relacionadas à Produção Intelectual seriam interessantes para a avaliação e tomada de decisão da universidade (ou ProPesq)?"*

Toda a conversa foi gravada e transcrita para posterior análise. Alguns trechos das entrevistas podem ser encontrados no Anexo 1.

4.4 Aplicação da técnica de Classificação de Termos

Utilizando a transcrição das entrevistas livres foi realizada uma análise de freqüência e relevância dos termos que apareceram durante as conversas. Naturalmente, termos comuns da língua portuguesa, como verbos de ligação e preposições, foram excluídos desta análise.

O resultado deste processo de análise das entrevistas foi um conjunto de 63 termos destacados, que foram transformados em fichas contendo um termo cada uma. A listagem dos termos a freqüência de cada um encontram-se no anexo 2.

O conjunto de fichas foi então utilizado em encontros com as mesmas pessoas entrevistadas anteriormente. A seqüência de classificações proposta era a seguinte:

1. Dividir o conjunto completo de termos em 2 ou 3 grupos, conforme preferência do especialista. Obter os nomes dos grupos. Observar novos termos e termos descartados.

2. Para cada grupo gerado no passo anterior, solicitar que o especialista divida-o em 2 ou 3 subgrupos. Obter os nomes destes subgrupos. Observar novos termos e termos descartados.

3. Repetir passos 1 e 2 enquanto o especialista se mostrar apto a realizar novas classificações e o tempo permitir.

Na SAI, a primeira divisão do conjunto completo resultou em dois grupos: (1) termos considerados descartáveis, com a grande maioria dos termos; e (2) chamado

Avaliação da Produção Intelectual da UFRGS, que para as especialistas era um resumo do instrumento que elas estavam desenvolvendo. Este segundo grupo foi disposto na forma de um diagrama, que refletiu exatamente a hierarquia das informações nos relatórios que elas estão elaborando. Estes relatórios são organizados da seguinte forma: uma seção para produção intelectual dos docentes e outra para os técnicos. Dentro de cada uma destas há três seções: produção bibliográfica, produção artística e produção técnica. Há três níveis de relatórios: para a universidade como um todo, para cada unidade e para cada departamento. O diagrama montado pode ser visualizado no anexo 2, sendo que o termo produção científica foi usado no lugar de produção bibliográfica.

Ao ser solicitado que o conjunto *Avaliação da Produção Intelectual da UFRGS* fosse subdivido, as especialistas afirmaram que não conseguiam ver outra forma de organizar os termos. Também foi sugerido que uma outra divisão fosse feita com o conjunto completo de termos, segundo outro critério, mas novamente as especialistas se recusaram. Quando inquiridas a respeito da ausência de termos referentes à estrutura organizacional da universidade, que haviam merecido destaque durante a entrevista, as especialistas resolveram refinar o diagrama, incluindo termos como *Departamento* e *Unidade*. Além disso adicionaram alguns termos representando detalhes sobre os repositórios de dados e sistemas de categorização da produção intelectual. O diagrama final obtido está apresentado na Figura 4.1. Mais detalhes podem ser encontrados no anexo 2.

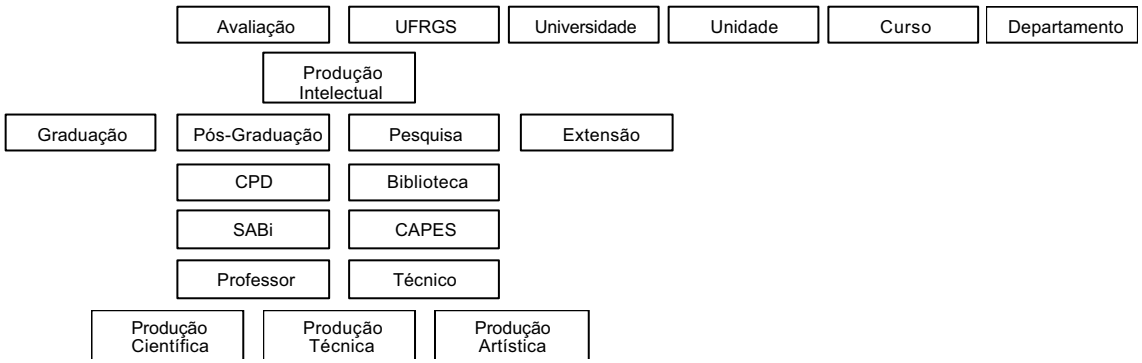


Figura 4.1 - Diagramação dos termos pela SAI

Depois disso, as especialistas declararam que não havia mais o que modificar, pois consideram que esta é a sua visão, já sedimentada pelo trabalho que vêm realizando.

No encontro na ProPesq a técnica foi mais produtiva: foi obtido um bom número de diferentes classificações para os termos e bastante informação oral surgiu durante o processo.

Apesar do escopo deste estudo de caso ter sido definido previamente focando na produção intelectual, na coleta de requisitos junto à ProPesq outros assuntos de interesse para a área da pesquisa universitária vieram à tona, como as questões sobre os agentes da pesquisa, a divulgação e impacto da pesquisa na sociedade e principalmente sobre os projetos de pesquisa, assunto recorrente para a vice pró-reitora. No anexo 2 há mais detalhes sobre estes tópicos. Nesta seção serão destacados apenas os resultados relacionados à produção intelectual.

Na primeira classificação de termos realizada na ProPesq, foram obtidos três grupos: (1) *Categorias de Produção Intelectual por Área de Conhecimento*; (2) *Agentes da Pesquisa*; e (3) termos descartados. O primeiro grupo evidenciou uma relação existente entre os objetos do domínio sobre a qual ela tem grande interesse, e que já havia sido destacada na entrevista. Trata-se da afinidade de certos tipos de produção intelectual com cada área do conhecimento. No anexo 2 está apresentado resultado desta divisão.

A divisão do grupo 1 gerou dois subgrupos: (1.1) *Produção Intelectual*; e (1.2) *Divisões do Lattes*[2]. O subgrupo 1.1 continha os termos que são identificados como categorias de produção, representando uma classe de objetos domínio, ou uma *dimensão*, na terminologia DW. O subgrupo 1.2 gerou discussões interessantes também, que podem ser observadas no anexo 2. A Tabela apresenta estes dois grupos.

Tabela 4.1 - Classificação de termos obtida na ProPesq

Produção Intelectual	Divisões do Lattes
Artigo Nacional	Biblioteca
Artigo Internacional	Avaliação
Livro	Formação
Patente	Lattes
Periódico	Projeto
Produção Artística	Publicação
Produção Bibliográfica	SABi
Produção Científica	Titulação
Produção Técnica	
Produto	
Produto Tecnológico	
Resumo	

O trabalho realizado com o grupo (2) *Agentes da Pesquisa* pode ser encontrado no anexo 2. Nele foram esclarecidas questões sobre o processo de pesquisa e formação de pesquisadores na universidade, que não estão diretamente relacionados ao tema de produção intelectual.

O passo seguinte foi realizar um segundo turno de classificações, agora com os termos dos grupos 1 e 2 anteriores misturados, excluindo-se os termos descartados (do grupo 3).

Como resultado da primeira classificação neste segundo turno foram obtidos dois grupos: um de descartados e outro chamado *Impacto dos Resultados da Pesquisa*, que é um tema importante para a ProPesq. Este segundo grupo foi então divido, gerando três subgrupos: *Iniciativas Internas de Divulgação*, *Iniciativas Externas de Divulgação* e *Produção Intelectual*. Novamente, este último grupo representa uma classe de termos, cuja característica comum é o fato de serem categorias de produção intelectual.

[2] Lattes é um sistema de formulários eletrônicos distribuído pelo CNPq para que os pesquisadores mantenham seus currículos atualizados em seu cadastro, informando, entre outras coisas, sua produção intelectual a atividades de pesquisa, por exemplo.

4.5 Elaboração do modelo de dados

Com o resultado do trabalho de coleta de requisitos foi possível compreender o domínio da atividade de pesquisa e especificamente da produção intelectual na universidade.

O tipo de modelo de dados utilizado foi o modelo dimensional, conforme apresentado no capítulo 2. As razões para escolha de modelo dimensional foram os benefícios destacados na literatura revisada: simplicidade, facilidade de navegação pelo usuário final e modelagem voltada aos indicadores usados na tomada de decisão e não na relação de dependência entre os dados.

Para a construção do modelo dimensional foram seguidos os passos discutidos na seção 2.3.1.4:

1. **Escolha do processo de negócio:** na verdade, este passo foi realizado antes mesmo do início do estudo de caso, no momento em que se decidiu que o assunto seria produção intelectual.

2. **Escolha da granularidade dos fatos:** a recomendação na literatura é de que se escolha a menor granularidade possível, desde que não cause explosão no espaço de armazenamento. Neste sentido a escolha é óbvia: cada registro na tabela de fatos será um produto intelectual apresentado à sociedade (publicado, implantado ou exibido, dependendo do tipo de produção – científica, técnica ou artística).

3. **Escolha das dimensões:** observando as transcrições das entrevistas e as classificações de termos foi possível identificar as dimensões pelas quais os usuários desejam acessar os dados de produção intelectual.

   Para a SAI a principal dimensão está relacionada com a autoria da produção intelectual. É importante saber dentro da estrutura organizacional da universidade para que órgão será contabilizada a produção, ou sejam dentro de que departamento ou unidade foi desenvolvido o produto intelectual. Esta informação está diretamente relacionada com importantes decisões tomadas pela universidade: é usada no cálculo dos índices para distribuição das vagas docentes entre os departamentos de ensino, por exemplo. Assim, a informação sobre o órgão ao qual está vinculada a produção intelectual é um dos atributos necessários.

   Também com relação à autoria da produção intelectual, a SAI necessita elaborar relatórios distintos para a produção dos técnicos e para a produção dos docentes. O tipo de vínculo é, portanto, outro atributo relevante sobre a autoria da produção intelectual.

   Ainda com relação à autoria, conforme relatado em entrevista, a ProPesq tem necessidade de tomar decisões que se baseiam na produção intelectual no nível individual do pesquisador, tais como avaliação de mérito para obtenção de financiamentos de projetos de pesquisa ou de bolsas para pesquisadores. Logo, é preciso saber quem é o autor de cada produto intelectual para que seja possível dar suporte a decisões como estas.

Com base nestas observações, a conclusão é que há uma dimensão **Autor** para a produção intelectual, com atributos identificando a pessoa, sua vinculação com a universidade (técnico ou docente) e o seu local de trabalho (departamento e unidade). Esta dimensão também poderia ser chamada *pesquisador*, o que seria natural para a produção intelectual da pesquisa, porém é preciso lembrar que há outros tipos de produção intelectual, como a artística, que nem sempre são resultado de pesquisa.

Uma dimensão que aparece claramente é a de **Categoria de Produção**. Tanto nas entrevistas quanto nas classificações obtidas sempre que se referencia produção intelectual é imediata a distinção entre produção bibliográfica ou (científica), técnica e artística. Estes são os três grandes tipos de produção e para cada um deles existe uma série de categorias de produção intelectual. Além disso, esta dimensão foi identificada mais de uma vez na forma de um conjunto de termos na ProPesq. Uma listagem desta categorização é encontrada no anexo 3. Estas categorias são importantes para o cálculo de índices departamentais de desempenho, pois de acordo com a categoria pode-se atribuir pontuações diferenciadas para a quantidade de produção do departamento. Não foi incluído um atributo para estas pontuações na dimensão porque esta é uma questão politicamente difícil e instável e dificilmente uma categoria de produção intelectual poderá ter uma pontuação definida valendo para toda a universidade.

A produção intelectual precisa ainda ser conhecida por **Área de Conhecimento**. As áreas de conhecimento são uma divisão do conhecimento humano proposta pelo CNPq. Uma das consultas que foi explicitamente citada na coleta de requisitos junto a ProPesq foi o cruzamento das Categoria de Produção com as Áreas de Conhecimento. É bastante interessante para a universidade traçar o perfil da forma como cada área de conhecimento se comunica com a sociedade através da sua produção intelectual. A definição destes perfis pode inclusive levar a uma solução para o problema de comparar a produção intelectual entre as áreas de conhecimento, missão muito difícil e que sempre gera descontentamento e divergências dentro da universidade, mas necessária para estabelecimento de índices de desempenho comparativo entre os departamentos, como os usados para a alocação de vagas em concursos públicos.

Por fim, como todo modelo dimensional, é preciso incluir a dimensão tempo, que nesse caso referencia-se ao **Ano**, visto que não foi identificada nenhuma necessidade de comparação no tempo em períodos menores que um ano. Isso se explica pelo fato de que a publicação de uma produção intelectual não é um evento que ocorra com grande freqüência, afinal é o resultado de um processo de pesquisa que em geral leva no mínimo alguns meses para ser concluído.

4. **Escolha dos fatos**: os fatos são geralmente valores numéricos associados ao grão da tabela de fatos. Mas neste caso não é possível encontrar fatos. Este é um caso de tabela de fatos sem fatos. O único fato importante é a existência do produto intelectual, que será utilizada então para contagens.

O modelo dimensional construído está apresentado na Figura . A notação utilizada para apresentação do modelo é de diagrama ER, mas o modelo é dimensional.

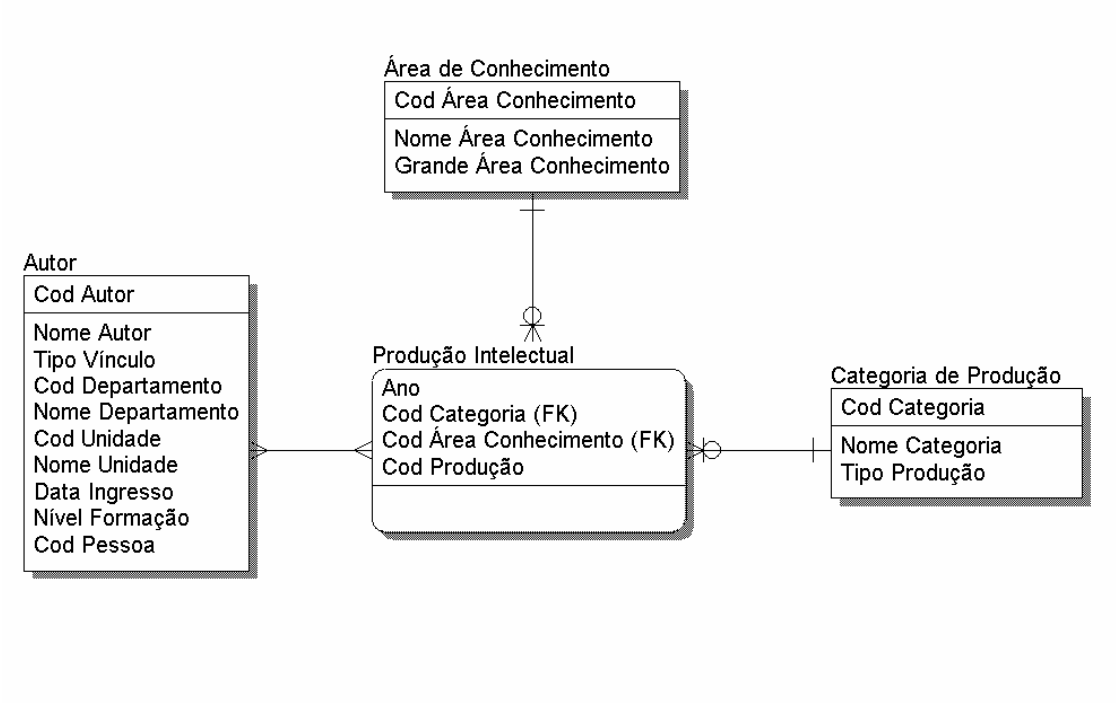


Figura 4.2 - Modelo dimensional para Produção Intelectual

A dimensão Ano é uma dimensão degenerada, pois não possui atributos. Por isso, ela está representada apenas pela presença da chave *Ano* na tabela de fatos.

A elaboração do modelo no esquema estrela na sua forma mais comum não foi possível, visto que o problema modelado tem características muito peculiares, distintas daquelas encontradas nos modelos tradicionalmente cobertos na literatura (varejo, empresas de seguros, etc.).

Uma peculiaridade neste modelo que o torna diferente dos modelos convencionais é a necessidade de um campo-chave na tabela de fatos que não é chave estrangeira (*Cod Produção*). Isto se deve ao simples fato de que um mesmo autor pode gerar mais de um produto intelectual de mesma categoria na mesma área de conhecimento no mesmo ano. Seria possível tentar resolver este problema utilizando uma data completa na dimensão tempo, porém, apesar de muito difícil, pode ocorrer duplicidade de chave ainda assim.

O relacionamento da dimensão Autor com a tabela de fatos é um relacionamento muitos-para-muitos e isso representa o maior desafio deste modelo. Um produto intelectual pode ter mais de um autor e esta é uma situação bastante comum, inclusive combinando autores de diferentes departamentos. Esta situação pode provocar uma dupla contagem de produtos, quando, por exemplo, solicitarmos a contagem da produção intelectual de um determinado departamento e dois docentes daquele departamento publicaram um artigo em conjunto. Uma forma de resolver isto é utilizando uma contagem apenas de itens distintos sobre o campo-chave da tabela de fatos (*Cod Produção*), opção comumente disponível em softwares de front-end (como

planilhas eletrônicas e pacotes estatísticos) e linguagens de consulta a bancos de dados relacionais (como SQL).

Outra característica do modelo que merece destaque é a dimensão Autor. Esta é uma dimensão que poderá sofrer alterações ao longo do tempo. Por exemplo, um docente ou técnico pesquisador poderá ser transferido de departamento. Neste caso, a opção mais adequada seria criar um novo registro para manter toda a sua produção anterior no departamento em que estava lotado. Para dar suporte a uma pesquisa sobre um autor e sua produção intelectual é necessário então um atributo que mantenha a relação entre registros de um mesmo autor que tenha sido transferido. Este é o propósito do atributo *Cod Pessoa*, que é a identificação única das pessoas na universidade no banco de dados operacional integrado mantido pelo CPD. Todos os registros de uma mesma pessoa na dimensão Autor devem utilizar o mesmo *Cod Pessoa*. Assim, para visualizar toda a produção intelectual de uma pessoa deverá ser feita a pesquisa por *Cod Pessoa*.

Uma descrição mais formal do modelo é encontrada no Anexo 5.

4.6 Levantamento de origens dos dados

A identificação dos repositórios de dados onde são encontrados os dados necessários para implementação do modelo proposto foi realizada internamente no CPD da universidade, visto que o autor deste trabalho é funcionário do referido centro e atua justamente na administração de dados, o que lhe confere o privilégio de conhecer todos os repositórios de dados institucionais mantidos pelo CPD. Todavia, informalmente foram realizados alguns esclarecimentos juntamente com os analistas de sistemas responsáveis pelas aplicações que suportam os repositórios de dados necessários.

Os dados referentes à dimensão Autor são facilmente obtidos no banco de dados administrativo (chamado de ADMINI), que mantém o cadastro de todos os servidores da universidade e seus locais de trabalho. É um banco de dados consistente e atualizado, sendo portanto uma fonte de dados qualificada.

A dimensão Categoria de Produção corresponde a uma tabela como a apresentada no anexo 3, que ainda deve ser completada, pois ainda está em definição pela SAI.

Da mesma forma, a dimensão Área de Conhecimento é também uma tabela, porém definida pelo CNPq, com quatro níveis hierárquicos: grande área de conhecimento, área de conhecimento, sub-área e especialidade. Para este trabalho bastam os dois maiores níveis. No Anexo 4 estão listadas as áreas de conhecimentos.

O grande problema é obter os dados para a tabela de fatos. A produção bibliográfica (científica) é com certeza encontrada em sua maioria no SABi, Sistema de Automação de Bibliotecas, que é o sistema "oficial" para o registro da produção intelectual. Porém este sistema apenas registra o que entra para o acervo da biblioteca (o que permite que se certifique a existência), e algumas categorias de produção técnica e artística não são passíveis de entrar para o acervo (por exemplo: desenvolvimento de fármaco ou apresentação de peça teatral). Existe também o currículo Lattes, distribuído pelo CNPq, que é utilizado pelos pesquisadores para informar seu currículo ao CNPq e

que é mais flexível para registrar a produção intelectual, mas que não impede que a pessoa informe produções não existentes. Uma vez que o CNPq considera o currículo Lattes dos pesquisadores no momento de conceder financiamentos, esta base de dados é freqüentemente atualizada pelos pesquisadores. Há uma proposta de realizar uma comparação entre o Lattes e o SABi, permitindo manter o SABi tão atualizado quanto o Lattes.

Mas mesmo que se consiga manter no SABi toda a produção intelectual da universidade, ela não estará registrada conforme o modelo proposto. A identificação dos autores não é compatível com o banco de dados administrativo mantido pelo CPD, a codificação usada para os assuntos nas bibliotecas não é compatível com as áreas de conhecimento do CNPq e a definição do tipo de item no acervo não é compatível com a categorização de produção intelectual proposta.

Para carregar os dados do SABi no DW será necessário um grande trabalho de integração com as dimensões. O escopo deste trabalho não inclui o processo de transformação de dados do back-end para o repositório de dados do DW, limitando-se a identificar os sistemas e repositórios de dados.

A Tabela 4 apresenta um resumo do mapeamento das origens de dados para as tabelas do modelo dimensional de Produção Intelectual proposto.

Tabela 4.2 - Origens dos dados para o modelo dimensional de Produção Intelectual

Sistema de Origem dos Dados	Tabela do Modelo Proposto
Banco ADMINI	Autor (dimensão)
Criação manual	Categoria de Produção (dimensão)
Criação manual	Área de Conhecimento (dimensão)
SABi	Produção Intelectual (fato)

5 Avaliação do Estudo de Caso

Um ponto determinante para o desenvolvimento do estudo de caso e que teve reflexos nos seus resultados foi a escolha do escopo do trabalho. A definição de um assunto tão focado acabou limitando a etapa de coleta de requisitos. A razão desta limitação vem do fato de que, ao escolher o assunto Produção Intelectual para o trabalho, foi também definido qual a atividade da universidade cujos resultados seriam incluídos no DW. Isso determinou que as únicas informações que estavam sendo levadas em consideração durante a coleta de requisitos eram aquelas relativas à produção intelectual. Ou seja, da forma como o assunto foi definido, a origem dos dados foi implicitamente determinada e as decisões e necessidades analíticas foram restringidas apenas àquelas que se valem apenas de dados da produção intelectual.

Na literatura sobre coleta de requisitos em DW, os autores [KIM98, POE99, KAC00] propõem que a coleta de requisitos seja realizada ou para a organização toda, ou para uma *área* de negócios, mas nunca para um *processo* do negócio. Os processos de negócio são escolhidos para a construção dos modelos dimensionais, cujos fatos estão diretamente relacionados com os resultados dos processos. O que ocorreu neste estudo de caso foi uma antecipação que acabou limitando a coleta de requisitos.

Retomando os objetivos da coleta de requisitos, vistos na seção 2.2, ao final do processo deve-se obter um amplo entendimento dos negócios do usuário participando da coleta. Ou seja, deve-se limitar o escopo da coleta de requisitos apenas pelos usuários que serão atendidos, e não pelos processos cujos resultados originam os dados usados nas medidas de desempenho dos negócios. Após conhecer os negócios do usuário, e identificar quais são as atividades que geram as informações desejadas, é que devem ser escolhidos os processos para os quais serão construídos modelos dimensionais.

Sobre o desenvolvimento da coleta de requisitos propriamente dita deve-se destacar as diferenças que ocorreram entre a SAI e a ProPesq. Durante todo o processo as analistas da SAI procuraram manter o foco e descrever apenas suas necessidades analíticas de informações da produção intelectual, necessidades estas que já estavam cristalizadas. Era preocupação constante das analistas da SAI que o foco não fosse desviado, para garantir que o projeto fosse concluído o quanto antes.

Na SAI não foi demonstrado interesse em incrementar o instrumento de avaliação que estava sendo desenvolvido. Em parte isso se deve à urgência em tê-lo implementado. Mas também é provável que as analistas da SAI realmente desconheçam outras formas de apresentar as informações sobre produção intelectual. Isso se refletiu no momento da aplicação da técnica de classificação de termos, quando as analistas da SAI reproduziram com os termos a visão já construída das informações e não foram capazes de rever seus conceitos

Na ProPesq os resultados foram enriquecidos pela multiplicidade de assuntos que acabaram sendo abordados. O estudo de caso junto à ProPesq acabou chegando mais próximo ao objetivo da coleta de requisitos em DW, ao permitir ao analista do projeto que tomasse conhecimento dos vários assuntos relevantes para a pesquisa universitária, entre os quais figura a produção intelectual. A “indisciplina” da vice pró-reitora de pesquisa ao não se manter presa ao assunto foi extremamente benéfica para os

resultados deste trabalho. Contribuíram para isto o fato de não haver restrições em função de prazos e o fato de que não existem relatórios já definidos, com longo de tempo de uso, que pudessem conter a criatividade da usuária[3].

É notável a diferença nos resultados obtidos com a aplicação da classificação de termos na ProPesq com relação aos obtidos na SAI. Isso certamente está relacionado com o perfil das pessoas participantes, suas expectativas e as demandas do seu departamento. A hipótese que emerge desta constatação é que o uso de técnicas de aquisição de conhecimento para a coleta de requisitos em DW é recomendável quando o usuário não tem ainda uma organização mental das informações suficientemente clara para conseguir expô-las verbalmente. De fato, a própria vice pró-reitora declarou que a técnica a ajudou a visualizar melhor o seu domínio de trabalho, o que está de acordo com o descrito na seção 3.2.2. E isto aconteceu sem a necessidade de algum talento especial da parte de quem está coletando os requisitos, o que reforça a qualidade das técnicas de aquisição de conhecimento em superar problemas de comunicação.

A construção do modelo de dados para produção intelectual demonstrou que o esquema estrela, amplamente utilizado em aplicações de DW pelo mundo, tem limitações importantes que devem ser contornadas artifícios. Estes artifícios podem prejudicar a compreensibilidade do modelo, um dos requisitos básicos de modelos de dados para suporte à decisão. Ainda há espaços a cobrir no estudo de modelagem dimensional, e sempre haverá, pois, assim como modelo de dados deste trabalho tem suas peculiaridades, muitos outros também terão.

Por fim, o levantamento de origens para os dados evidenciou que os dados no ambiente operacional podem ser encontrados em formato inadequado para o suporte à decisão. Isso reafirma a necessidade da coleta de requisitos ser realizada antes da construção do DW, pois evita um esforço que possivelmente não traria benefícios por não estar em conformidade com as expectativas dos usuários.

[3] A ProPesq é uma pró-reitoria relativamente jovem na universidade e ainda não tem uma cultura de relatórios de análise

6 Conclusões

Os executivos e gerentes de uma organização em alguns casos tomam suas decisões com base em questões subjetivas e pessoais, como sua experiência e seu *feeling,* ao invés de se basear em fatos objetivos. Nesta situação, o executivo dificilmente será capaz de relatar como toma decisões e que informações são usadas no processo, pois ele não tem uma organização mental das informações. Neste caso, o uso de técnicas de aquisição de conhecimento para a coleta de requisitos de um DW será uma opção atraente.

Este trabalho apresentou um estudo sobre o uso de técnicas de aquisição de conhecimento na coleta de requisitos e modelagem de dados para DW.

Foram revisados conceitos importantes sobre data warehousing e enfatizadas as características de um DW como sistema de apoio a decisão. As técnicas utilizadas para a coleta de requisitos e modelagem de dados usadas tradicionalmente em DW foram discutidas.

O processo de tomada de decisão organizacional baseado em DW foi revisado no contexto da Engenharia de Conhecimento, procurando demonstrar as contribuições que esta disciplina pode dar para a área de DW através do uso das técnicas de Aquisição de Conhecimento para a coleta de requisitos de DW.

Por fim, um estudo de caso envolvendo modelagem de dados foi desenvolvido e avaliado para analisar as vantagens obtidas com a aplicação de técnicas de aquisição de conhecimento na coleta de requisitos em DW.

Sobre o uso de técnicas de aquisição de conhecimento na coleta de requisitos este estudo permitiu concluir que seu uso pode trazer benefícios, especialmente em casos onde é difícil para o usuário do DW descrever verbalmente quais informações são usadas no processo de tomada de decisão ou análise do negócio.

A técnica de classificação de termos, utilizada no estudo de caso deste trabalho, é um exemplo de técnica que possui uma característica particularmente desejável para situações como a descrita acima: ela pode levar o especialista visualizar uma organização mental das informações que ele mesmo utiliza mas que nunca havia articulado conscientemente antes.

Adicionalmente, algumas conclusões não relacionadas diretamente com o objetivo podem ser descritas.

O escopo da coleta de requisitos para DW deve ser definido com base no setor da organização que se deseja contemplar, selecionando-se os usuários adequados e não os processos que originam os dados.

O esquema estrela tem limitações que precisam ser contornadas artifícios que podem prejudicar a compreensibilidade do modelo, eliminando as vantagens de um modelo simples. A modelagem dimensional ainda precisa ser confrontada com domínio mais complexos do que os tradicionais para amadurecer.

A necessidade da coleta de requisitos ser realizada antes da construção do DW foi reforçada ao demonstrar-se que os dados no ambiente operacional não atendiam aos requisitos impostos pelos usuários.

Como sugestão para estudos futuros podem ser citados o uso de outras técnicas de aquisição de conhecimento na coleta de requisitos de DW, mais experimentos como este para comprovar se há ou não vantagens e mais estudos sobre como transformar o conhecimento obtido num modelo de dados adequado para suporte à decisão.

Outra possibilidade de estudos futuros está ligada à crescente interação entre data warehousing e engenharia de conhecimento que vem sendo promovida pela Gestão do Conhecimento. Propostas recentes [KER01, NEM02] sugerem que o data warehouse evolua para o knowledge warehouse, através do uso de bases de conhecimento integradas ao data warehouse. O uso de técnicas de aquisição de conhecimento na coleta de requisitos para o DW neste contexto também poderia ser usada para preencher estas bases de conhecimento.

7 Referências

[ABE01] ABEL, M. **Estudo da Perícia em Petrografia Sedimentar e sua Importância para Engenharia de Conhecimento**. Tese de Doutorado. Porto Alegre, CPGCC da UFRGS. 2001

[ABE02] ABEL, M. **Sistemas de Conhecimento**. Notas de Aula. Porto Alegre, Instituto de Informática da UFRGS. 2002.

[BAR97] BARQUIN, R., EDELSTEIN, H. **Planning and Designing the Data Warehouse**. New Jersey, Prentice Hall. 1997.

[DEM01] DEMAREST, M. **Data Legibility in Decision Support Systems (DSS)**. White Paper. Decision Points Applications Inc. Oregon. 2001.

[GOL98] GOLFARELLI, M., MAIO, D., RIZZI, S. **The Dimensional Fact Model: a Conceptual Model for Data Warehouses**. Invited Paper, International Journal of Cooperative Information Systems, vol. 7, n. 2&3, 1998. Disponível por FTP em ftp://ftp-db.deis.unibo.it/pub/stefano/ijcis98.pdf. Data da última visita: 28/02/2003.

[GON99] GONÇALVES, V. P. **Modelagem de dados para DataWarehouse**. Trabalho de diplomação. Porto Alegre, Instituto de Informática da UFRGS. 1999.

[HEU01] HEUSER, C. A. **Projeto de Banco de Dados**. Porto Alegre, Sagra-Luzzato. 2001.

[INM96] INMON, W. H. **Building the Data Warehouse**. New York, John Wiley & Sons. 1996.

[KAC00] KACHUR, R. J. **The Data Warehouse Diary: Requirements Gathering for Data Warehouse Design**. DM Review. 2000. Disponível por WWW em http://www.dmreview.com. Data da última visita: 28/02/2003.

[KER01] KERSCHBERG, L. **Knowledge Management in Heterogeneous Data Warehouse Environments**. In International Conference on Data Warehousing and Knowledge Discovery. Munich, Germany. 2001

[KIM96] KIMBALL, R. **The Data Warehouse Toolkit**. New York, John Wiley & Sons. 1996.

[KIM98] KIMBALL, R. et al. **The Data Warehouse Lifecycle Toolkit**. New York, John Wiley & Sons. 1998.

[LAM95] LAMBERT, Bob. **Break Old Habits to Define Data Warehouse Requirements**. DM Review. Disponível por WWW em http://www.dmreview.com. Data da última visita: 28/02/2003.

[MOR98] MORAES, R. L. **Sistemas de Data Warehouse: Estudo e Aplicação na Área da Saúde**. Porto Alegre, CPGCC da UFRGS. 1998.

[NEM02] NEMATI, H. R. et al. **Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing**. Decision Support Systems, vol. 33 issue 2, pp143-161. Elsevier, 2002.

[POE98] POE, V., KLAUER, P., BROBST, S. **Building a Data Warehouse for Decision Support**. New Jersey, Prentice Hall. 1998.

[RAD96] RADEN, N. **Modeling the Data Warehouse**. Information Week, Issue 564, janeiro 1996. Disponível por WWW em http://user.aol.com/nraden/iw0196_1.htm. Data da última visita: 28/02/2003.

[SCH99] SCHREIBER, G. et al. **Knowledge Engineering and Management: the CommonKADS Methodology**. Cambridge, The MIT Press. 1999.

[SHI02] SCHIM, J. P. et al. **Past, present, and future of decision support technology**. Decision Support Systems, vol. 33 issue 2, pp111-126. Elsevier, 2002.

Anexo 1 – Trechos de entrevistas para coleta de requisitos

Entrevista na Secretaria de Avaliação Institucional (SAI)

L: Em termos de produção intelectual, quais são as informações que a Universidade precisa para tomar decisões?
A: Nós pegamos a categorização de produção intelectual da CAPES, embora nesta categorização não tenha atuação internacional, que se refere a professores pesquisadores que atuem como professores visitantes em instituições estrangeiras, por exemplo, isto aqui tem que entrar como produção intelectual porque faz parte deste conjunto. Depois vem produção bibliográfica, produção artística, produção técnica e as teses, dissertações, monografias e trabalhos de conclusão e patentes registradas. Então acho que pegaríamos estas cinco subcategorias do que podemos entender por produção intelectual.
A: Então, por que que nós pegamos a categorização da CAPES? Porque ela traz uma discriminação, um tipo de produção que os docentes nas diferentes áreas da universidade fazem. O que acontecia antes: se considerava mais a produção bibliográfica e toda a produção artística, por exemplo, ficava num limbo. Acho que depois poderíamos falar sobre a experiência da alocação das vagas docentes.
E a produção técnica ficava num meio termo, entre produção intelectual e produção não acadêmica. E a CAPES agora tem uma discriminação tão bem refinada que te dá a possibilidade de localizar qualquer tipo de produção dentro de uma destas categorias. Então nós resolvemos adotar. Houve um trabalho aqui dentro da universidade, para a alocação de vagas docentes nos departamentos, onde não se usou o modelo CAPES e a Marlis pode te relatar a dificuldade que foi poder garantir que a produção artística, por exemplo, fosse considerada no mesmo nível que produção intelectual.
L: A CAPES então também definiu uma série de atributos que qualificam uma produção intelectual como um tipo ou outro.
A: Sim, na verdade, colocaram assim: a produção intelectual está organizada em termos de produção bibliográfica, artística e técnica. Então tu pega todas as áreas. Por exemplo, quando é um produto mais tecnológico, o resultado de uma pesquisa que redunda num produto tecnológico, não é produção intelectual no sentido restrito do termo, então por isso até eu proponho esta outra categoria que registra as patentes, que é uma coisa mais recente que a universidade começa a ir atrás.

Entrevista na Pró-Reitoria de Pesquisa (ProPesq)

L: Da produção intelectual, o que é utilizado para a tomada de decisões na ProPesq?
M: temos um programa de fomento, como é que a gente avalia se uma pessoa vai ganhar um auxílio ou não vai ganhar. Temos um programa de bolsas, que critérios se usou para dizer quem tem prioridade 1, 2, 3, desfavorável. Todas estas coisas é claro que evidentemente a pró-reitoria não decide isso, decide com os pesquisadores, porque na pesquisa todas as coisas são avaliadas pelos pares. E essa é a legitima avaliação dentro da área de pesquisa. Se eu tenho um periódico que vai ser avaliado: vale a pena a UFRGS investir neste periódico, nesta área? Um consultor ad-hoc da área vai dizer: isto aqui tem mérito. Então a pró-reitoria vai buscar os recursos. Então todos os programas tem uma comissão assessora de professores, mas esses professores têm uma orientação da pró-reitoria, tem formulários propostos pela pró-reitoria com itens que eles devem observar e se manifestar. Então com relação a tudo entram sempre as mesmas coisas... básicas, e às vezes nas situações particulares acrescidas de outras. Mas as básicas estão ligadas à produção e não só ao âmbito do projeto, à importância do projeto, claro que isso conta. Em primeiro lugar o que se olha: o currículo do professor. Hoje, fala-se em currículo fala-se em Lattes. Então lá no Lattes: o que que ele fez. E aí tem uma confusão que a gente tem que Ter cuidado. Porque a gente tem oito áreas pelo CNPq. Cada área tem o seu jeito de se comunicar com a comunidade científica. Vamos pegar a Informática, por exemplo. A Informática tem uma forma muito peculiar que é a apresentação em congresso de resumo expandido. Essa é uma comunicação forte. Se eu falar na Física, o que que um consultor olha no currículo de um professor da Física? No mínimo o cara tem que Ter um artigo internacional, num periódico de circulação internacional. Bom mas aí se eu vou para a Agronomia, não é esta mesma visão. O consultor vai ver quantas publicações em revistas nacionais ele teve? Vale mais a nacional. Se eu for para a Bioquímica, para as biológicas, etc.: é um misto. Ainda é um misto. Mas todo mundo tem que olhar o que? O que que vale? Publicações. E aí entramos na questão: qualquer publicação? Em qualquer periódico? Não. Aí entra a questão assim: este periódico é o mais qualificado da área? Porque ele vale quanto mais pontos quanto for qualificado. Realmente então a pró-reitoria para tomar uma decisão ela vai lá e pega o currículo de uma pessoa e conta quantas publicações nacionais internacionais ela tem? Não. Mas alguém faz por ela. Então na verdade a pró-reitoria faz. Então tu vai dizer assim: só quem tem um número grande de pontos, os pesquisadores sêniors são as pessoas para quem a pró-reitoria aprova coisas? Absolutamente. A pró-reitoria sabe os patamares, sabe o que tá no patamar Ter um doutorando com dois anos de recém-doutor. Que vai ser diferente para um professor sênior. Então realmente a produção é um item decisivo.

Anexo 2 – Termos utilizados e Classificações obtidas durante eliciação de conhecimento

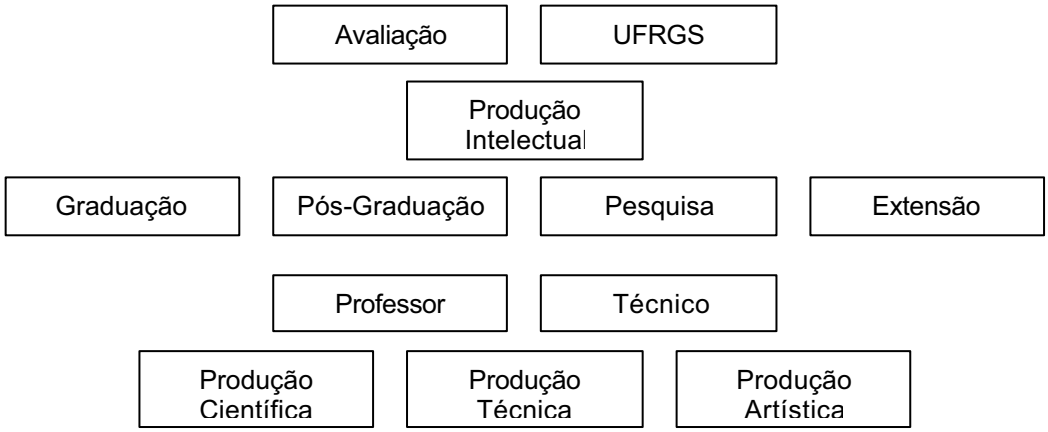
Termos utilizados para classificação

A listagem dos termos destacados a partir das entrevistas com suas correspondentes freqüências de ocorrência está apresentada na tabela a seguir:

Termo	Freqüência	Termo	Freqüência
Aluno	16	Patente	3
Ano	23	Periódico	5
Área	41	Pesquisa	33
Area de Conhecimento	4	Pesquisador	12
Artigo	5	Pós-graduação	17
Artigo Internacional	4	Processo	7
Avaliação	14	Produção Artística	7
Biblioteca	4	Produção Bibliográfica	5
Bolsista	3	Produção Científica	4
CAPES	6	Produção Intelectual	33
Categoria	6	Produção Técnica	3
CNPq	13	Produto	5
Comissão	4	Produto Tecnológico	3
Comissão Pesquisa	3	Professor	54
Currículo	10	Programa de Pós-Graduação	7
Currículo de Professor	4	Projeto	32
Curso	8	Pró-reitoria	16
Departamento	43	Publicação	24
Discriminação	5	Resumo	4
Dissertação	4	SABi	6
Doutor	6	Substituto	5
Extensão	22	Técnico	13
Formação	4	Tese	7
Graduação	7	Titulação	4
Grupo	22	Trabalho	15
Grupo de Pesquisa	6	Trabalho Conclusão	3
Iniciação Científica	5	UFRGS	9
Lattes	9	Unidade	24
Livro	11	Universidade	27
Mestrado	3	Vaga	6
Monografia	3	Visitante	8
Orientador	4		

Classificações obtidas na Secretaria de Avaliação Institucional

As especialistas da SAI descartaram a quase totalidade dos cartões logo de início e montaram a seguinte estrutura com os termos não descartados:

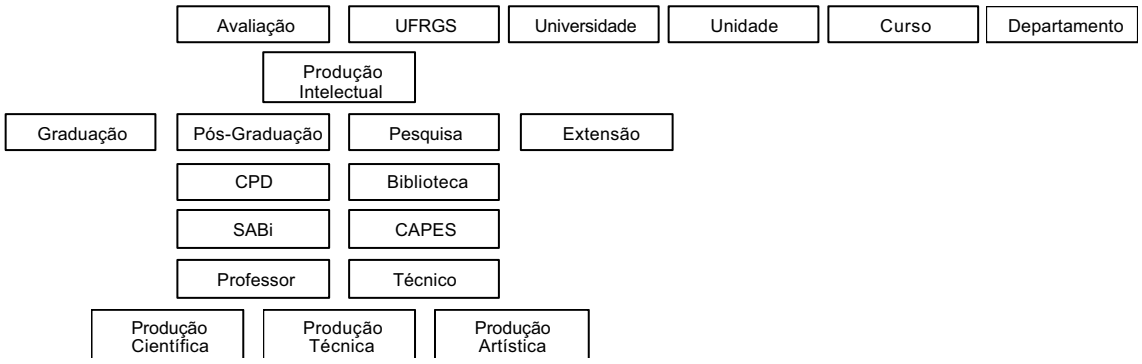


Na parte de baixo seriam incluídas as várias categorias de produção para cada um dos tipos (científica, técnica e artística). Professor e técnico são os autores da produção intelectual. As três fichas de cima são o nome da estrutura montada: Avaliação da Produção Intelectual na UFRGS.

Percebe-se nesta atitude das especialistas a preocupação em manter o escopo do estudo bem reduzido, para acelerar a conclusão do projeto e obter os dados necessários o quanto antes.

Quando solicitado que uma nova classificação fosse realizada, utilizando algum outro critério, as especialistas afirmaram que não havia outra possibilidade na visão delas. Isto certamente se deve ao fato de que a SAI já tem definido exatamente os dados que deseja e não está interessada em rever suas definições.

Quando inquiridas a respeito da ausência dos aspectos organizacionais da universidade (unidades e departamentos), que haviam sido amplamente destacados durante a entrevista, as especialistas argumentaram que para elas o termo UFRGS englobaria tal estrutura. Porém elas resolveram refinar a estrutura montada, gerando uma segunda versão, mais completa, apresentada abaixo:



CAPES e SABi representam aqui os sistemas de categorização da produção intelectual utilizados por cada um. CPD e Biblioteca são os locais onde são armazenados os dados. Expansão horizontal da ficha da UFRGS representa a estrutura organizacional da universidade, de forma hierárquica. Curso tem pouca importância.

Classificações obtidas na Pró-Reitoria de Pesquisa

Na primeira divisão foram obtidos três grupos, cada um representado por uma coluna da tabela abaixo, cujos cabeçalhos são os nomes dos grupos.

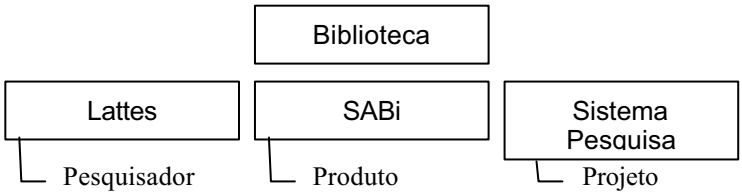
Categorias de Produção Intelectual por Área de Conhecimento	Agentes da Pesquisa	Descartados
Área de Conhecimento	Bolsista	Aluno
Artigo Nacional	CAPES	Ano
Artigo Internacional	CNPq	Área
Avaliação	Comissão Pesquisa	Comissão
Biblioteca	Dissertação	Currículo
Categoria	Doutor	Currículo de Professor
Formação	Grupo de Pesquisa	Curso
Lattes	Iniciação Científica	Departamento
Livro	Orientador	Discriminação
Patente	Pesquisa	Extensão
Periódico	Pesquisador	Graduação
Produção Artística	Pós-Graduação	Grupo
Produção Bibliográfica	Pró-Reitoria	Mestrado
Produção Científica	Processo	Monografia
Produção Intelectual	Técnico	Professor
Produção Técnica	Tese	Programa de Pós-Graduação
Produto		Substituto
Produto Tecnológico		Trabalho
Projeto		Trabalho de Conclusão
Publicação		UFRGS
Resumo		Unidade
SABi		Universidade
Titulação		Vaga
		Visitante

Os termos destacados em negrito no primeiro grupo foram identificados como sendo o nome do grupo pela especialista.

Logo após esta divisão foi solicitado à especialista que dividisse o primeiro grupo novamente. O resultado obtido foi:

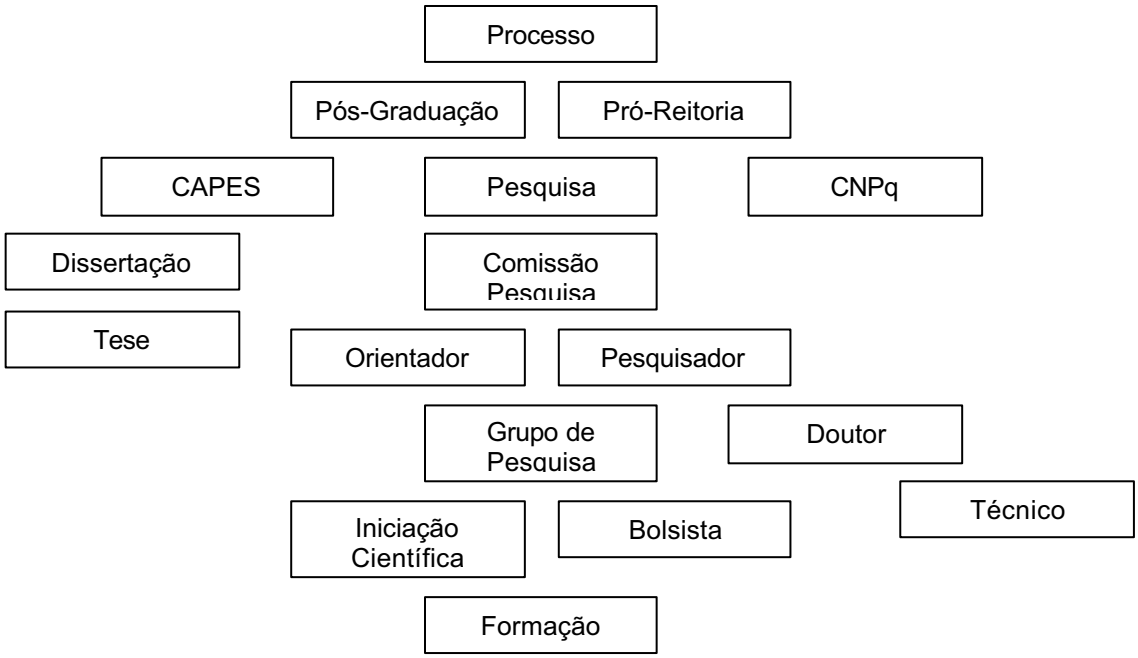
Produção Intelectual	Divisões do Lattes
Artigo Nacional	Biblioteca
Artigo Internacional	Avaliação
Livro	Formação
Patente	Lattes
Periódico	Projeto
Produção Artística	Publicação
Produção Bibliográfica	SABi
Produção Científica	Titulação
Produção Técnica	
Produto	
Produto Tecnológico	
Resumo	

A partir da discussão sobre o grupo “Divisões do Lattes” a especialista estruturou algumas das fichas que de certa forma sintetizam a sua visão sobre os dados relevantes para a pró-reitoria e os locais onde estão armazenados. A estrutura criada está apresentada a seguir:



Observando a figura percebe-se três grandes assuntos para a pesquisa: o pesquisador, cujos dados estão principalmente no currículo Lattes; a produção intelectual, com dados no SABi; e os projetos de pesquisa, com dados no Sistema Pesquisa, desenvolvido pelo CPD da UFRGS.

A próxima atividade foi estudar o grupo “Agentes da Pesquisa”, definido na primeira classificação. Desta vez, ao invés de subdividir o grupo a especialista optou por diagramar os termos. A estruturação obtida está apresentada a seguir:



Aqui a especialista retratou a sua visão sobre como se dá o processo de pesquisa na universidade e a formação de pesquisadores, demonstrando a relação entre os agentes da pesquisa. No alto se vê as duas pró-reitorias que alavancam a pesquisa: a pró-reitoria de pós-graduação e a pró-reitoria de pesquisa (representada apenas por pró-reitoria). CAPES e CNPq são as principais agências de fomento e foram colocadas ao lado da pró-reitoria com a qual mais se relacionam institucionalmente. As teses e dissertações estão relacionadas à pós-graduação e por isso ficaram no braço esquerdo do diagrama. As comissões de pesquisa são também responsáveis por alavancar a pesquisa dentro das unidades e são elas que mantém contato mais direto com o pesquisador. Pesquisador e orientador são dois papéis desempenhados pelos docentes e por isso estão lado a lado. Um pesquisador em geral é um doutor e também pode ser um técnico. Por fim, os grupos de pesquisa reúnem os pesquisadores/orientadores e seus bolsistas/orientandos, sejam eles de Iniciação Científica ou de outro tipo, e é nesta relação dentro do grupo de pesquisa que se dá a formação de novos pesquisadores.

Esgotada a discussão sobre o diagrama, foi realizado um segundo turno de classificações, agora misturando-se todas as fichas que não foram descartadas na primeira classificação.

O resultado na primeira classificação deste segundo turno foi o seguinte:

Impacto dos Resultados da Pesquisa	Descartados
Artigo Nacional	Área de Conhecimento
Artigo Internacional	Avaliação
CAPES	Biblioteca
CNPq	Categoria
Comissão Pesquisa	Dissertação
Grupo de Pesquisa	Doutor
Lattes	Formação
Livro	Iniciação Científica
Patente	Orientador
Periódico	Pesquisa
Pós-Graduação	Pesquisador
Pró-Reitoria	Processo
Produção Artística	Técnico
Produção Bibliográfica	Tese
Produção Científica	Titulação
Produção Intelectual	
Produção Técnica	
Produto	
Produto Tecnológico	
Projeto	
Publicação	
Resumo	
SABi	
Sistema Pesquisa	

Com as fichas que não foram descartadas foi solicitado que a especialista realizasse uma subdivisão, que resultou na seguinte classificação:

Iniciativas Externas de Divulgação	Iniciativas Internas de Divulgação	Produção Intelectual
CAPES	Comissão Pesquisa	Artigo Nacional
CNPq	Pós-Graduação	Artigo Internacional
Grupo de Pesquisa	Pró-Reitoria de Pesquisa	Livro
Lattes	Sistema Pesquisa	Patente
		Periódico
		Produção Artística
		Produção Bibliográfica
		Produção Científica
		Produção Intelectual
		Produção Técnica
		Produto
		Produto Tecnológico
		Projeto
		Publicação
		Resumo
		SABi

Anexo 3 – Categorização de Produção Intelectual

Tipo de Produção	Categoria de Produção
Produção Bibliográfica	Artigo publicado em periódico especializado de circulação local: trabalho completo
	Artigo publicado em periódico especializado de circulação local: resumo
	Artigo publicado em periódico especializado de circulação nacional: trabalho completo
	Artigo publicado em periódico especializado de circulação nacional: resumo
	Artigo publicado em periódico especializado de circulação internacional: trabalho completo
	Artigo publicado em periódico especializado de circulação internacional: resumo
	Livro: capítulo
	Livro: coletânea
	Livro: texto integral
	Livro: verbete
	Livro: outro
	Trabalho em anais: trabalho completo
	Trabalho em anais: resumo
	Tradução: artigo
	Tradução: livro
	Tradução: outro
	Partitura musical: canto
	Partitura musical: coral
	Partitura musical: orquestral
	Partitura musical: outro
	Artigo em jornal ou revista comuns
Produção Artística	Apresentação de obra artística: coreográfica
	Apresentação de obra artística: literária
	Apresentação de obra artística: musical
	Apresentação de obra artística: outra
	Arranjo musical: canto
	Arranjo musical: coral
	Arranjo musical: orquestral
	Arranjo musical: outro
	Composição musical: canto
	Composição musical: coral
	Composição musical: orquestral
	Composição musical: outro
	Programa de rádio e televisão: dança
	Programa de rádio e televisão: música
	Programa de rádio e televisão: teatro
	Programa de rádio e televisão: outro
	Obra de artes visuais: cinema
	Obra de artes visuais: desenho
	Obra de artes visuais: escultura
	Obra de artes visuais: fotografia
	Obra de artes visuais: gravura
	Obra de artes visuais: instalação
	Obra de artes visuais: pintura
	Obra de artes visuais: televisão
	Obra de artes visuais: vídeo
	Obra de artes visuais: outra
	Sonoplastia: cinema

Produção Artística	Sonoplastia: música
	Sonoplastia: rádio
	Sonoplastia: teatro
	Sonoplastia: televisão
	Sonoplastia: outra
Produção técnica	Apresentação de trabalho: comunicação
	Apresentação de trabalho: conferência
	Apresentação de trabalho: congresso
	Apresentação de trabalho: seminário
	Apresentação de trabalho: simpósio
	Apresentação de trabalho: outra
	Carta, mapa, similar: aerofotograma
	Carta, mapa, similar: carta
	Carta, mapa, similar: fotograma
	Carta, mapa, similar: mapa
	Carta, mapa, similar: outro
	Desenvolvimento de aplicativo: computacional
	Desenvolvimento de aplicativo: multimídia
	Desenvolvimento de aplicativo: outro
	Desenvolvimento de material didático
	Desenvolvimento de material instrucional
	Desenvolvimento de produto: aparelho
	Desenvolvimento de produto: instrumento
	Desenvolvimento de produto: equipamentos fármacos e similares
	Desenvolvimento de produto: outro
	Desenvolvimento de técnica: analítica
	Desenvolvimento de técnica: instrumental
	Desenvolvimento de técnica: pedagógica
	Desenvolvimento de técnica: processual
	Desenvolvimento de técnica: terapêutica
	Desenvolvimento de técnica: outra
	Editoria: edição
	Editoria: editoração
	Editoria: outra
	Manutenção de obra artística: arquitetura
	Manutenção de obra artística: desenho
	Manutenção de obra artística: escultura
	Manutenção de obra artística: fotografia
	Manutenção de obra artística: gravura
	Manutenção de obra artística: pintura
	Manutenção de obra artística: outra
	Maquete
	Produção de programa de rádio e televisão: entrevista
	Produção de programa de rádio e televisão: mesa redonda
	Produção de programa de rádio e televisão: comentário
	Produção de programa de rádio e televisão: outro
	Relatório final de pesquisa

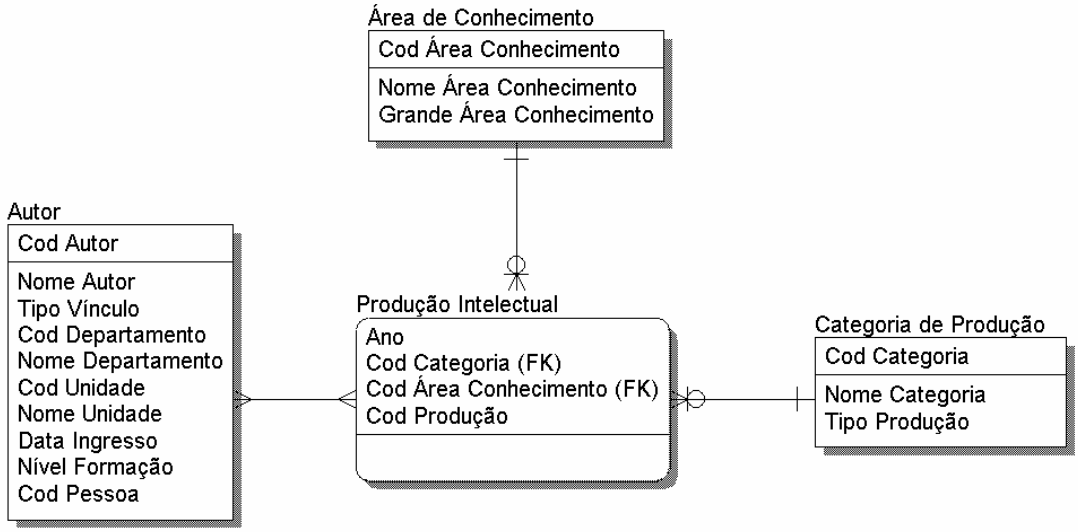
Anexo 4 – Listagem das Áreas de Conhecimento

Grande Área do Conhecimento	Área do Conhecimento
Ciências Agrárias	Agronomia
	Ciência e Tecnologia de Alimentos
	Engenharia Agrícola
	Medicina Veterinária
	Recursos Florestais e Engenharia Florestal
	Recursos Pesqueiros e Engenharia de Pesca
	Zootecnia
Ciências Biológicas	Biofísica
	Biologia Geral
	Bioquímica
	Botânica
	Ecologia
	Farmacologia
	Fisiologia
	Genética
	Imunologia
	Microbiologia
	Morfologia
	Parasitologia
	Zoologia
Ciências da Saúde	Educação Física
	Enfermagem
	Farmácia
	Fisioterapia e Terapia Ocupacional
	Fonoaudiologia
	Medicina
	Nutrição
	Odontologia
	Saúde Coletiva
Ciências Exatas e da Terra	Astronomia
	Física
	Geociências
	Matemática
	Oceanografia
	Probabilidade e Estatística
	Química
Ciências Humanas	Antropologia
	Arqueologia
	Ciência Política
	Educação
	Filosofia
	Geografia
	História
	Psicologia
	Sociologia
	Teologia

Ciências Sociais Aplicadas	Administração
	Arquitetura e Urbanismo
	Ciência da Informação
	Comunicação
	Demografia
	Direito
	Economia
	Economia Doméstica
	Museologia
	Planejamento Urbano e Regional
	Serviço Social
	Turismo
Engenharias e Ciência da Computação	Ciência da Computação
	Desenho Industrial
	Engenharia Aeroespacial
	Engenharia Biomédica
	Engenharia Civil
	Engenharia de Materiais e Metalúrgica
	Engenharia de Minas
	Engenharia de Produção
	Engenharia de Transportes
	Engenharia Elétrica
	Engenharia Mecânica
	Engenharia Naval e Oceânica
	Engenharia Nuclear
	Engenharia Química
	Engenharia Sanitária
Lingüística, Letras e Artes	Artes
	Letras
	Lingüística

Anexo 5 – Descrição do Modelo de Dados para Produção Intelectual

O modelo de dados (dimensional) gerado no estudo de caso está apresentado logo abaixo:



As descrições de cada tabela são apresentadas a seguir:

Tabela: Autor
Tipo: Dimensão
Descrição: Representa o conjunto de autores de produção intelectual.
Atributos:

Cod Autor	Chave primária gerada automaticamente (sem significado)
Nome Autor	Nome do Autor
Tipo Vínculo	Indica se o Autor é docente ou técnico
Cod Departamento	Código do Departamento conforme banco de dados de origem (ADMINI)
Nome Departamento	Nome do Departamento
Cod Unidade	Código da Unidade conforme banco de dados de origem (ADMINI)
Data Ingresso	Data em que o autor ingressou na universidade
Nível Formação	Nível de formação do autor (graduação, especialização, mestrado, doutorado)
Cod Pessoa	Identificação única do autor (para mapear diferentes registros de Autor da mesma pessoa)

Tabela: Área de Conhecimento
Tipo: Dimensão
Descrição: Tabela de áreas de conhecimento definida pelo CNPq
Atributos:

Cod Área Conhecimento	Chave primária gerada automaticamente (sem significado)
Nome Área Conhecimento	Nome da área de conhecimento
Grande Área Conhecimento	Nome da grande área de conhecimento na qual a área de conhecimento está contida

Tabela: Categoria de Produção

Tipo: Dimensão
Descrição: Representa a categorização de produção intelectual definida pela SAI.
Atributos:

Cod Categoria	Chave primária gerada automaticamente (sem significado)
Nome Categoria	Nome da categoria de produção intelectual
Tipo Produção	Tipo de produção intelectual (bibliográfica, artística ou técnica)

Tabela: Produção Intelectual
Tipo: Fato
Descrição: Registra cada produto intelectual gerado pela universidade
Atributos:

Ano	Ano do produto intelectual (dimensão degenerada)
Cod Categoria	Chave estrangeira para dimensão Categoria de Produção
Cod Área Conhecimento	Chave estrangeira para dimensão Área de Conhecimento
Cod Produção	Chave gerada automaticamente para cada produto intelectual (sem significado)