SECRETARIA ESPECIAL DE POLÍTICAS PARA AS MULHERES

ONU MULHERES

Sistema de Informação da SPM

Produto 6 - Mecanismo de coleta/alimentação de dados - ETL

Jaqueline Juvencio de Sá

Produto 6 – Elaboração de mecanismo de coleta/alimentação de dados a partir da definição de indicadores baseados em informações produzidas pelas áreas técnicas da SPM.

Contrato n. 28/2015

Objeto da contratação: Aperfeiçoamento de aplicativos eletrônicos de gestão da informação relacionados à implementação e acompanhamento de políticas, apoiando a implementação de procedimentos e mecanismos que aumentem o potencial de uso de ferramentas de gestão de informações no processo de coordenação e articulação relacionados à SPM.

Valor do produto: R$ 10.400,00 (Dez mil e quatrocentos reais)

Data de entrega: 12/07/2016

Nome do consultor: Jaqueline Juvencio de Sá

Nome do supervisor: Filipe Hagen E. da Silva

De Sá, Jaqueline Juvencio

Título do produto: Mecanismo de coleta/alimentação de dados – ETL/2016.

Total de folhas: 18

Supervisor: Filipe Hagen E. da Silva

Secretaria Especial de Políticas para as Mulheres

Palavras-chave: Extração, transformação, carga.

SUMÁRIO

RESUMO

Após a definição das bases de dados utilizadas no sistema de informação da SPM e a necessidade de consolidá-las em um banco de dados interno, foi criado um processo de extração, transformação e carga desses dados. Esse é um passo muito importante do projeto, visto que os novos dados serão atualizados no *Data Warehouse* por meio dessa ferramenta e por isso deve ser de fácil manutenção, permitindo que os processos possam ser alterados conforme necessário.

Palavras-Chave: Extração, transformação, carga.

1. INTRODUÇÃO

1.1 Esta consultoria terá como objetivo geral apoiar a SPM no aperfeiçoamento de aplicativos eletrônicos de gestão da informação relacionados à implementação e acompanhamento de políticas, apoiando a implementação de procedimentos e mecanismos que aumentem o potencial de uso de ferramentas de gestão de informações no processo de coordenação e articulação relacionados à SPM. Este Produto específico tem como objetivo definir o processo de extração, transformação e carga das bases selecionadas para o sistema de informação da SPM.

1.2 Contexto e importância do Produto: As bases de dados selecionadas para esse projeto são estruturadas e organizadas de diferentes formas, dificultando a integração entre elas. Para que as bases sejam tratadas corretamente e gerem informações úteis, esse produto tem como objetivo específico criar processos para que os dados brutos sejam convertidos de maneira correta e mantenham a integridade dos dados, bem como facilitar o processo de limpeza e armazenamento dos dados.

2. DESENVOLVIMENTO

Nas etapas anteriores desse projeto, foram selecionadas as fontes de dados que alimentariam as ferramentas do sistema de informação da SPM. À medida que as ferramentas foram sendo criadas, a equipe do Observatório Brasil da Igualdade de Gênero foi selecionando as bases com a finalidade de formar um quadro mínimo de indicadores, a fim de atender as necessidades da Secretaria no sentido de proporcionar o acesso facilitado e o acompanhamento dos dados para ações de planejamento e execução das políticas públicas para as mulheres.

As bases selecionadas foram:

- IBGE Censo Demográfico Domicílio;
- IBGE Censo Demográfico Pessoas;
- IBGE PNAD Domicílio;
- IBGE PNAD Pessoas;
- INEP Censo Superior Aluno;
- INEP Censo Superior Curso;
- INEP Censo Superior Docente;
- INEP Censo Superior IES;
- MS SIM;
- MS SINAM/VIVA
- MTE RAIS;
- SPM Conselhos;
- SPM Disque 180 - Crimes Associados;
- SPM Disque 180 - Telefonia;
- SPM Disque 180 – Ouvidoria;
- SPM OPM;

- TSE Candidatos;
- TSE Eleitorado;
- TSE Votação Municipal.

Nesta etapa, será realizada a criação do processo de ETL (Extração, Transformação e Carga) para os dados adquiridos na construção do *Data Warehouse*. O ETL (do acrônimo inglês *Extract, Transform, Load*) é o processo de extração dos dados de fontes externas, o tratamento por meio de algoritmo de limpeza e armazenamento no *Data Warehouse*.

O processo de ETL exige uma grande parte do tempo de construção de um projeto como esse. Serão extraídos dados de fontes heterogêneas e o *Data Warehouse (DW)* precisa receber dados de forma homogênea e concisa para que sejam geradas informações de apoio à decisão e não apresentem resultados errôneos. A estrutura do processo de ETL foi desenhada de acordo com as informações relevantes e as necessidades de cada base que será extraída de vários sistemas de informação.

Como podemos perceber, esse processo possui três etapas. A primeira é a extração, a segunda, transformação e limpeza dos dados e, por fim, a carga para o *DW.*

2.1 Processo de Extração

Nessa parte é determinada a origem dos dados, seus metadados, tamanhos de cada variável de acordo com os valores disponíveis no dicionário de variáveis. Em geral, as bases utilizadas nesse projeto possuem formatos do tipo Arquivo Texto comum (.TXT), Banco de dados Dbase (.DBF), e Colunas Separadas por vírgula (.CSV). A extração converte as bases para um formato uniforme, que depende da ferramenta utilizada no processo, para a entrada no processo de transformação.

2.2 Processo de Transformação

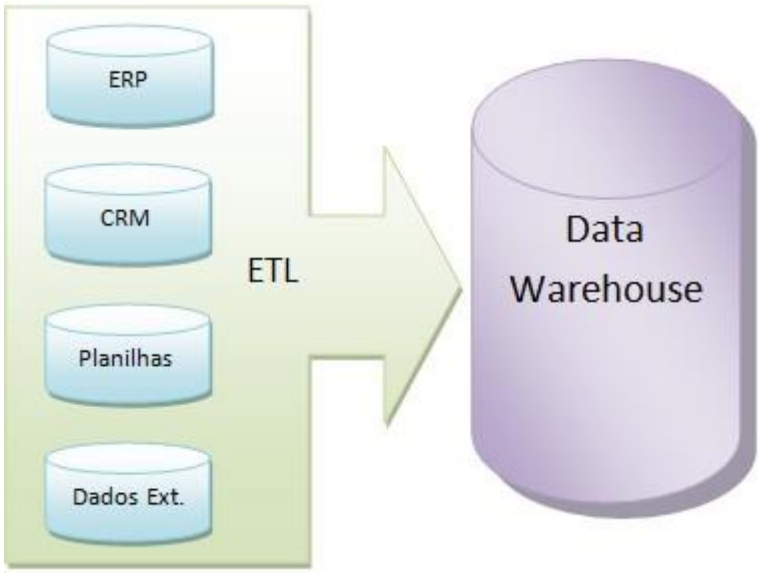
Nesse estágio pode ser aplicada uma série de regras ou funções para que os dados sejam carregados de forma homogênea. A maioria das fontes selecionadas nesse projeto não necessitou de muita manipulação de dados, pois seguem padrões parecidos.

Também é necessária a limpeza dos dados, que consiste em padronizar os dados de forma que permita absorver as variações no formato dos dados de entrada. Por exemplo, em um sistema a definição do formato de data é DD-MM-AAAA, em outras apresentam a mesma data no formato AAAA-MM-DD. Para que as informações de diferentes bases possam ser comparadas, esses formatos precisam ser compatíveis.

2.3 Processo de Carga

Após a ferramenta ETL fazer a parte da extração dos dados das várias fontes selecionadas, tratá-los por meio das definições feitas a partir de análises detalhadas dos dados de origem, será realizado processo de carga no DW para que sejam lidos e analisados. O tempo necessário para que o processo de carga possa ser realizado por completo varia de acordo com o volume de dados de cada base.

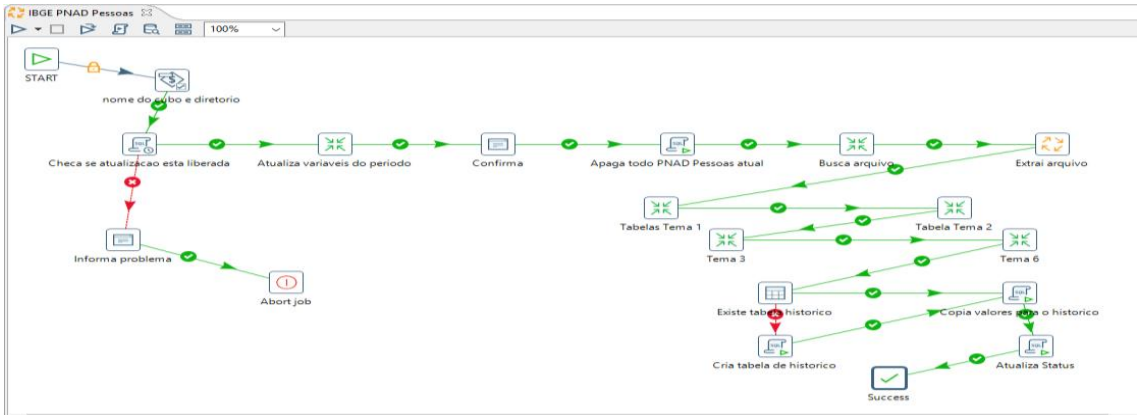
Figura 1: Processo ETL



2.4 Ferramenta utilizada para o processo de *ETL*

Para realizar essas etapas, a ferramenta escolhida foi a *Pentaho Data Integration -PDI ou Spoon*, produzida pela *Pentaho* e distribuída gratuitamente, com código aberto. Essa ferramenta tem ampla aceitação no mercado. Os *ETL* preparados para a SPM foram organizados de acordo com os Cubos criados anteriormente. Os *ETLs* são formados por dois tipos de arquivos: a) os *Jobs* (.kjb) contêm a descrição geral do processo de extração de cada base; b) as *Transformações* (.ktr) realizam a leitura dos arquivos de entrada, selecionam os

Figura 2: Job inicial do ETL

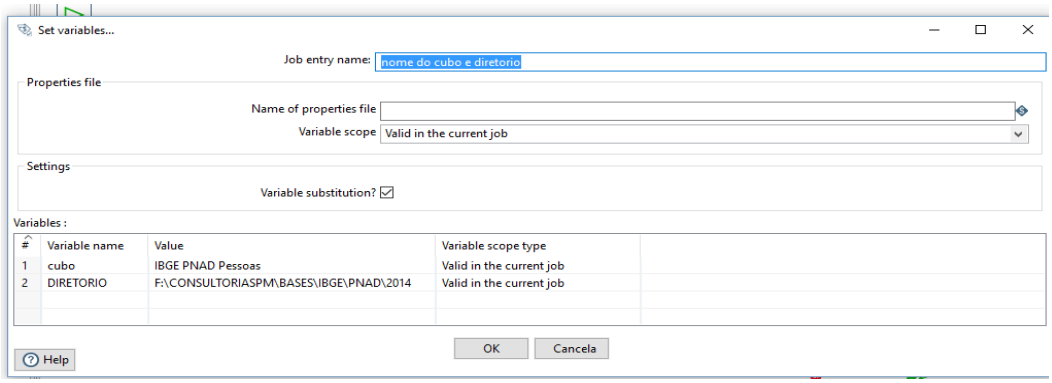


campos de interesse, realizam o processo de transformação quando necessário e armazenam a informação no banco. Todos os processos de *ETL* criados para as bases selecionadas possuem basicamente os mesmos componentes iniciais, como no exemplo abaixo:

Os componentes são os seguintes

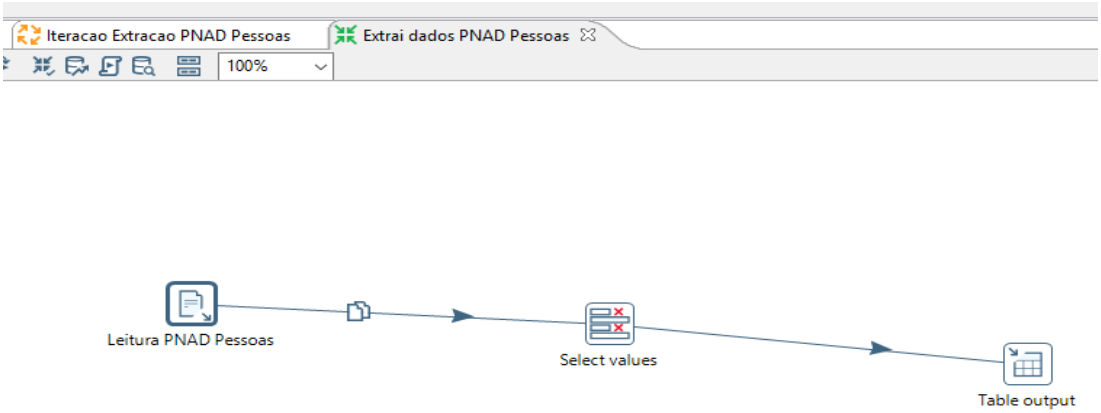
- "**START**": Início do processo de ETL.
- **"Nome do cubo e diretório**": Nele é possível indicar o nome do cubo que será atualizado e o caminho onde a base estará disponível.

Figura 3: Componente Set Variables



- **"Checa se atualização está liberada**": Faz uma consulta ao banco para verificar se a base foi liberada para atualização e qual o mês e ano que foi liberada.
- "**Informa problema**": Se a base não foi liberada, informa o problema e finaliza o *Job.*
- "**Atualiza variável do período**": Salva os novos valores de mês e ano nas variáveis atuais.
- "**Confirma**": Informa o mês/ano e o diretório informados no início do processo e pergunta se quer continuar.

Figura 4: Processo de Extração, transformação e carga.



- "**Apaga tabela atual**": Apaga os valores da última atualização da base.
- "**Busca Arquivos**": Caso a base possua mais de um arquivo, o componente busca esses arquivos e envia para o próximo componente.
- "**Extrai arquivo**": Um dos principais componentes do processo de ETL. Nele são realizadas as tarefas de transformação e carga dos dados. A *Spoon* possui uma ampla coleção de componentes pré-fabricados, capazes de processar vários formatos de arquivos de entrada.

Os próximos componentes são os de atualização das tabelas que alimentam os gráficos do Painel de Visualização de indicadores. Nem todas as bases possuem essa função.

- "**Existe tabela histórico**": Verifica no banco de dados se existe uma tabela de histórico criada para a base. Caso não exista, passa para o próximo componente "**Cria tabela histórico**" que possui o *script* de criação da tabela histórico.
- "**Copia valores para o histórico**": Possui um *script* para selecionar os valores da tabela atualizada e carrega na tabela histórico.
- "**Atualiza Status**": Atualiza os valores de configuração do processo de *ETL* da base e bloqueia novo processo até a próxima atualização.

A ferramenta *PDI* possui outros componentes que auxiliam em várias outras necessidades de transformação das bases. Para mais informações sobre esses componentes podem ser encontradas na documentação disponibilizada pela *Pentaho* (http://www.pentaho.com/).

2.5 Ferramenta administrativa do *ETL*

Como instrumento auxiliar do processo *ETL*, foi customizada a ferramenta que controla quais *ETLs* podem ser executados e quando, reduzindo chance de que atualizações que não estejam prontas para serem executadas sejam disparadas por acidente.

Na construção da ferramenta administrativa foi utilizado um *framework* de segurança para aplicações java, chamado *Apache Shiro*. O seu funcionamento depende de outras ferramentas customizadas nessa consultoria, como o Painel de visualização dos indicadores (Painel Observa Gênero) e o Cubo. A parte do layout está definida no painel e a visualização

dos cubos para liberação do ETL está definida no arquivo de configuração do Cubo. Mais informações estarão no manual de manutenção e no manual de utilização do sistema, que estará disponível no produto final dessa consultoria.

Essa ferramenta é uma aplicação Java Web, o que significa dizer que ela é executada em um servidor de aplicações java e a interação com o usuário é feita pelo navegador de internet. A seguir serão apresentadas telas do sistema.

Figura 4: Tela de login

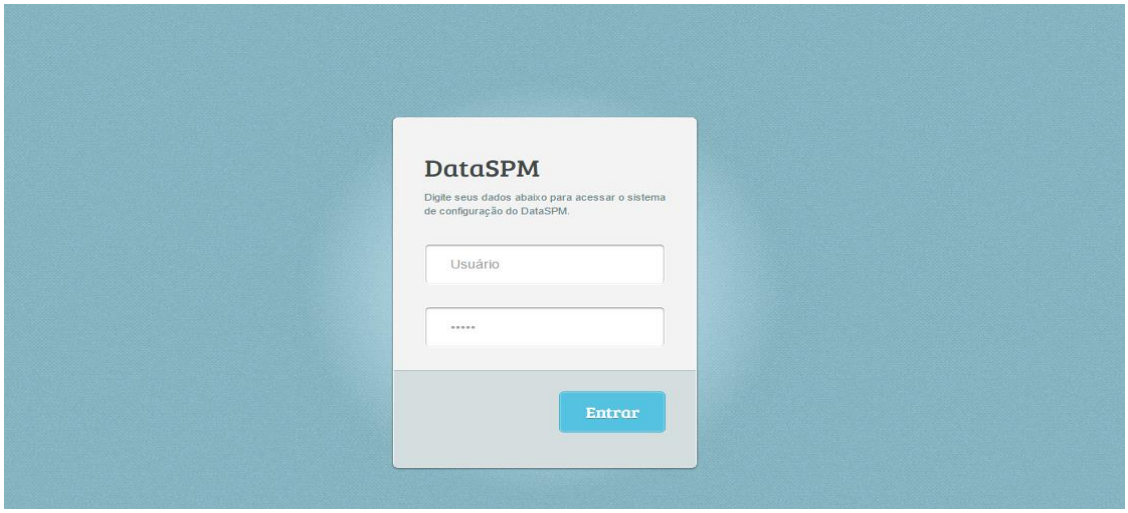


Figura 5: Tela inicial com as informações dos cubos

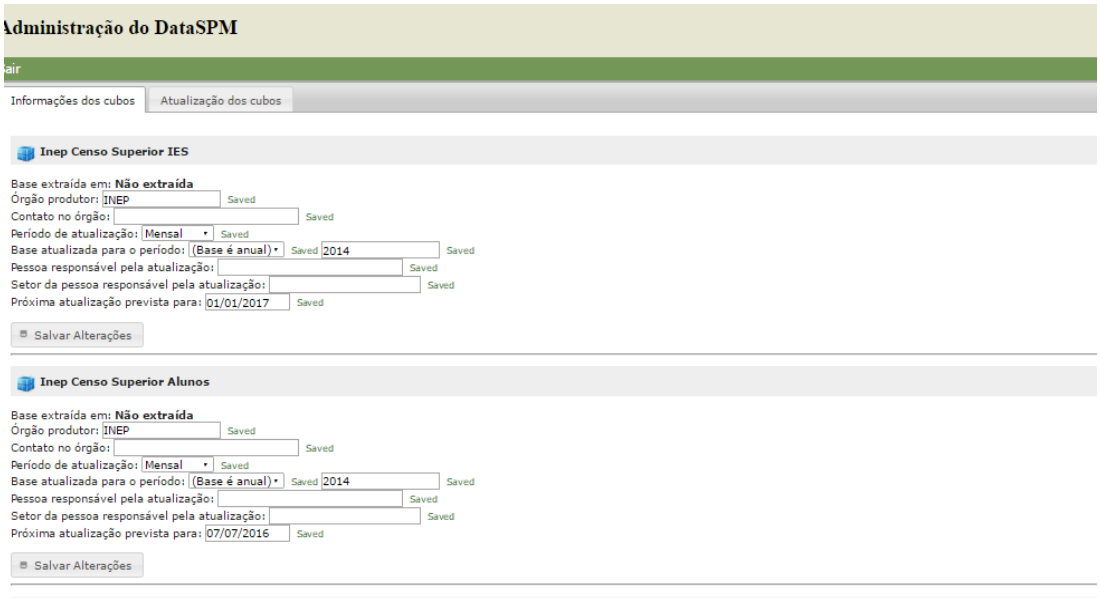
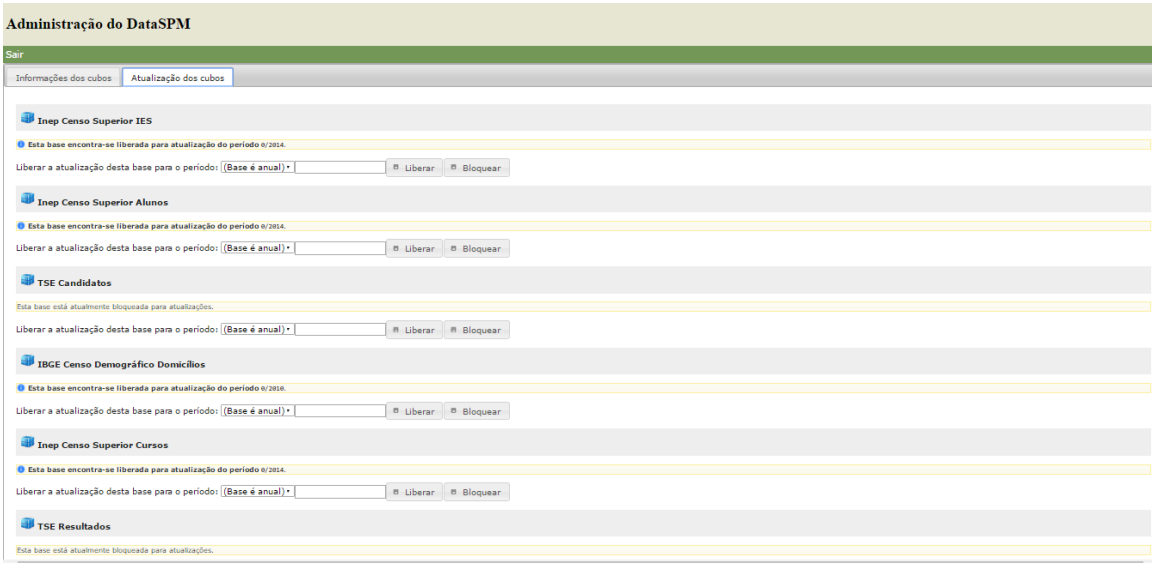


Figura 6: Tela para liberação dos processos de ETL



2.6 Informações sobre os processos de ETL da SPM

Os processos de *ETL* construídos nessa consultoria são específicos para as bases de dados selecionadas, assim, as bases precisam seguir um formato padrão que seja mantido consistentemente pelo órgão produtor. Na maioria dos casos as bases utilizadas seguem esse padrão. Porém, em alguns casos, a base pode mudar a versão acrescentando ou eliminando algumas variáveis. Como o processo de *ETL* não pode identificar essas mudanças automaticamente, é necessário que eles sejam reprogramados.

O ETL é um processo que permite dados brutos – de pesquisas socioeconômicas, de registros administrativos, entre outros – se tornarem informações que melhorem a gestão das políticas públicas. Assim, é necessário o envolvimento de um ou mais agentes que desenvolva o papel de “Gerente de dados”. Ele tem o objetivo de obter dados de órgãos produtores, validá-los (efetuando ajustes quando necessário) e inseri-los no banco de dados, por meio do

ETL. O “Gerente de dados” precisa ter conhecimento sobre os dados e sobre os processos necessários para se obtê-los.

Na tabela a seguir serão apresentadas informações sobre a periodicidade das bases que fazem parte do sistema de informação da SPM até o momento.

Tabela 1 - Atualização das bases

Base	**Período**	**Atualização do ETL**
IBGE Censo Demográfico Domicílio	Decenal	Decenal
IBGE Censo Demográfico Pessoas	Decenal	Decenal
IBGE PNAD Domicílio	Anual	Anual
IBGE PNAD Pessoas	Anual	Anual
INEP Censo Superior Aluno	Anual	Anual
INEP Censo Superior Cursos	Anual	Anual
INEP Censo Superior Docente	Anual	Anual
INEP Censo Superior IES	Anual	Anual
MS SIM	Contínuo	Anual
MS SINAN/VIVA	Contínuo	Anual
MTE RAIS	Anual	Anual
SPM Conselhos*	Semestral ou Anual	Semestral ou Anual
SPM Disque 180 Crimes Associados*	Semestral ou Anual	Semestral ou Anual

SPM Disque 180 – Telefonia*	Semestral ou Anual	Semestral ou Anual
SPM Disque 180 – Ouvidoria*	Semestral ou Anual	Semestral ou Anual
SPM OPM*	Semestral ou Anual	Semestral ou Anual
TSE Candidatos	Bienal	Bienal
TSE Eleitorado	Bienal	Bienal
TSE Votação Municipal	Bienal	Bienal

Nota: A periodicidade de atualização das bases destacadas (*) serão definidas a critério da direção da SPM.

3. CONCLUSÃO

Esse é o produto principal para união dos dados oriundos de sistemas externos com o sistema de informação da SPM. Os processos foram planejados e desenhados respeitando as necessidades de cada base, incluindo, também, os requisitos que foram levantados durante todo o projeto de criação do sistema de informação da SPM. É importante que os processos sejam executados de maneira correta, observando os períodos de atualização de cada base e principalmente, por um agente que tenha conhecimento técnico para administrar as bases.

ANEXOS

1. DVD com os arquivos de definição do *ETL.*