

DOI: 10.5748/9788599693124-13CONTECSI/RF-4219

ETL PROCESS USING OPEN DATA FROM E-GOV FOR DISCOVERY AND KNOWLEDGE REPRESENTATION

Leandro Quingerski (Universidade Federal de Santa Catarina, Santa Catarina, Brazil) - leandroquinger@gmail.com

Demis Marques (Universidade Federal de Santa Catarina, Santa Catarina, Brazil) - demismarq@gmail.com

Patricia de Sá Freire (Universidade Federal de Santa Catarina, Santa Catarina, Brazil) - patriciadesafreire@gmail.com

ABSTRACT: Data management, information and knowledge generated or acquired by organizations is a growing challenge today, no different with government organizations, whose mandatory provision of data brought by the Open Data Policies allows citizens to request data from any public organ, since the provided data or information is not in some way classified as confidential. Even considering the various existing initiatives, public data still depend on treatment for use and integration with other data, even if they are also from the government. The treatment of these data and information is important either for its recovery, interrelationships or yet to knowledge discovery since the information or data although the significant value without treatment, correlation or suitable connection may have little meaning for the users of this data or information. This paper aims to present a process proposal for ETL (extract, transform, load) process using open government data source for discovery and knowledge representation.

KEYWORDS: e-government, ETL, knowledge representation

RESUMO: O gerenciamento de dados, informações e conhecimentos gerados ou adquiridos pelas organizações é um desafio crescente nos dias atuais, não sendo diferente com as organizações governamentais, cuja obrigatoriedade de disponibilização dos dados trazida pela Lei de Acesso à Informação permite ao cidadão solicitar dados de qualquer órgão público, desde que esse dado ou informação não esteja de alguma maneira classificado como sigiloso. Mesmo considerando as diversas iniciativas existentes, os dados públicos dependem ainda de tratamento para utilização e integração com outros dados, mesmo que também sejam do governo. O tratamento destes dados e informações é importante seja para sua recuperação, inter-relação ou ainda para descoberta de conhecimento, uma vez que a informação ou os dados embora tenham significativo valor, sem um tratamento, correlação ou contexto adequado pode ter pouco significado para o usuário deste dado ou informação. Este trabalho tem como objetivo apresentar uma proposta de processo ETL (Extract, Transform, Load) utilizando dados abertos de fonte governamental para descoberta e representação de conhecimentos.

PALAVRAS-CHAVE: governo eletrônico, ETL, representação de conhecimento

1. INTRODUÇÃO

O gerenciamento de dados, informações e conhecimentos gerados ou adquiridos pelas organizações é um desafio crescente nos dias atuais, em especial com relação aos diferentes formatos e usos dos dados e informações a partir de necessidades específicas de uma organização.

O avanço dos sistemas corporativos e a produção acelerada de dados e informações usando diferentes tecnologias trazem o desafio aos gestores a problemática de como integrar e relacionar o conteúdo gerado transformando em conhecimento.

Essa problemática não é diferente quando se trata de dados e informações públicas originárias de organismos do governo, cuja obrigatoriedade de disponibilização dos dados trazida pela Lei de Acesso a Informação permite ao cidadão solicitar dados de qualquer órgão público, desde que esse dado ou informação não esteja de alguma maneira classificado como sigiloso.

A variabilidade de formatos e atividades geradoras de dados no serviço público pode inclusive intensificar a dificuldade de interoperabilidade. Considerando isso, algumas iniciativas como a utilização de metadados padronizados, o e-PMG (Padrão de Metadados para Governo Eletrônico, e a INDA (Infraestrutura Nacional de Dados Abertos, ou ainda a INDE (Infraestrutura Nacional de Dados Espaciais), pretendem orientar a um conjunto de normativas direcionadas a padronização de dados públicos de acordo com a tipologia do dado e informação e seu respectivo uso (Governo Eletrônico, 2015).

Ações mais abrangentes como o e-PING (Padrão de Interoperabilidade de Governo Eletrônico) pretendem definir um conjunto de diretrizes para interação entre os diferentes Poderes e a sociedade em geral (e-PING, 2015).

Algumas outras iniciativas dizem respeito ao uso de ontologias para integração e interoperação dos dados e informações, exemplo disso é o Modelo Ontológico da Classificação da Despesa do Orçamento Federal Brasileiro, e a Ontologia de Participação Social, ambas desenvolvidas pelo Governo Federal Brasileiro (Governo Eletrônico, 2015).

Mesmo considerando as diversas iniciativas existentes, os dados públicos dependem ainda de tratamento para utilização e integração com outros dados de fontes governamentais.

O tratamento destes dados e informações é importante seja para sua recuperação, inter-relação ou ainda descoberta de conhecimento a partir de relações existentes ou novas relações entre dos dados e informações, uma vez que a informação ou o dados embora tenha significativo valor, sem um tratamento, correlação ou contexto adequado pode ter pouco significado para o usuário deste dado ou informação, e pode assim representar pouco em relação ao objetivo inicial de captura deste dado ou informação, onde o dado ou informação não gera ou representa um conhecimento significativo para o usuário.

Considerando esta problemática relacionada a necessidade de tratamento de dados para uso e mais especificamente para descoberta e representação do conhecimento este trabalho tem como objetivo apresentar uma proposta de processo ETL (Extract, Transform, Load) utilizando dados abertos de fontes governamentais para descoberta e representação de conhecimentos.

Para isso foi realizado levantamento na literatura de um conjunto de métodos e técnicas que pudessem colaborar para a elaboração de um processo para realização deste trabalho. Posteriormente foram feitos diversos testes utilizando bases de dados e informações governamentais, com um conjunto de ferramentas pré-selecionadas para que se pudesse chegar a um processo cujos resultados fossem expressivos quanto ao que se objetivava alcançar. Por fim foram definidos metadados, em um perfil de metadados, para integração e operação dos dados e informações coletadas.

Como resultado foi possível elaborar um processo ETL utilizando dados governamentais abertos, onde foram padronizados e processados utilizando um conjunto de tarefas e ferramentas, resultando num processo capaz de integrar dados e informações, gerando novas informações e conhecimentos, além de permitir diferentes representações onde os dados e informações da fonte original não permitiam.

Desta maneira foi possível apresentar uma proposta de processo de ETL utilizando dados abertos de fontes governamentais demonstrando que este tipo de processo, mesmo que simplificado, pode potencializar a utilização de dados governamentais abertos para gerar um conjunto novo de informações e conhecimentos a partir da padronização e relacionamento entre dados e informações.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são tratados os principais conceitos envolvidos neste trabalho, cabe salientar que a intenção não é exaurir estes conceitos, sendo que já existem várias literaturas que tratam sobre os mesmos. O objetivo é tratar de forma sucinta os assuntos abordados neste trabalho a fim de contextualizar os próximos capítulos.

2.1. Governo eletrônico (e-gov)

Governo eletrônico pode ser entendido como uma infraestrutura de rede compartilhada por diferentes órgãos públicos a partir da qual a gestão dos serviços públicos é realizada (Rover, 2008). A partir da otimização desses serviços o atendimento ao cidadão são realizados, visando atingir a sua universalidade, bem como ampliando a transparência das suas ações e permitindo o controle social.

O governo eletrônico tem forte potencial para aprimorar a prestação de serviços, melhorar a interação entre indivíduos, empresas e instituições civis, e ao mesmo tempo, reduzir os custos e tempo associados à prestação de serviços públicos. Além disso, pode fornecer maior autonomia e capacidade de participação dos cidadãos, permitindo a todos contribuírem diretamente no processo de governança pública, bem como ser um catalisador para o desenvolvimento econômico e social (Rotta, 2013).

Algumas práticas associadas ao Governo Eletrônico podem ser vistas como a transição de uma gestão pública fortemente hierarquizada e burocrática para uma reorganização do Estado num formato horizontal, colaborativo, flexível e inovador, alinhado ao conceito de sociedade do conhecimento (Rotta, 2013).

Cabe ressaltar, que a consolidação e integração de políticas de serviços em todos os níveis e agentes do governo também fazem parte do contexto que engloba o conceito de governo eletrônico.

Para Koh e Prybutok (2003), ações de governo eletrônico seriam impossíveis sem a sustentação fornecida por modernas tecnologias de informação e comunicação, as quais apoiam na criação e disponibilização de novos canais e ambientes de interação e comunicação. O elemento de conectividade de TIC fornece canais de informação e feedback entre as agências governamentais e as partes interessadas.

As tecnologias de TIC permitem a rápida transmissão de informações e dados, entre os agentes que estiverem conectados na rede do governo eletrônico, aumentando assim, a qualidade e quantidade de informações recebidas (Rover, 2008).

2.2. Processo de ETL

O conceito de processo de ETL (*Extract, Transform and Load*) está relacionado à extração, transformação e carga de dados de uma ou mais fontes de origem para uma ou mais bases de dados de destino (Abreu, 2008).

O processo de extração e carga, são obrigatórios no processo, consistindo respectivamente na obtenção dos dados de uma fonte primária e na carga uma base de destino (Kimball, 1998). A transformação é necessária quando se precisa trabalhar de alguma maneira os dados obtidos, antes da realização da carga dos mesmos em um banco de dados.

Na literatura geralmente o conceito de ETL está associado a criação de um data warehouse (DW), porém o conceito vai além da construção deste tipo de base de dados, e se estende também para qualquer processo em que se necessite trabalhar com dados desde a extração, transformação e carga dos mesmos.

2.3. Representação e descobertas de conhecimentos

O conhecimento permeia todas as atividades organizacionais e é utilizado por todos os membros de uma organização. Neste contexto, também estão inseridas as organizações governamentais (Grant, 2005).

Como um recurso fundamental para a operação do governo, o conhecimento pode gerar insumos para alcançar resultados da ação política e pública. E com a evolução da área de engenharia do conhecimento, uma série de metodologias e técnicas são desenvolvidas e aperfeiçoadas, aumentando as possibilidades em trabalhar com a gestão do conhecimento dentro de uma organização pública.

Como a Engenharia do Conhecimento se refere a todos os aspectos técnicos, científicos e sociais envolvidos na construção, manutenção e uso de Sistemas Baseados em Conhecimento, torna-se evidente a importância que esta área possui na construção de um governo eletrônico eficiente (Rautenberg, 2009)

A Gestão do Conhecimento é compreendida como um fator essencial ao processo de produção, contribuindo com melhorias nos resultados, processos e na qualidade dos produtos desenvolvidos, podendo promover maior efetividade para uma organização e conseqüentemente para os envolvidos.

Com o conhecimento sendo visto como processo, entende-se que é possível fazer a gestão do mesmo. Diversos autores separam o processo de gestão do conhecimento (GC) em quatro etapas, sendo: Criação, Armazenamento e Recuperação, Transferência e Aplicação (Alavi & Leidner, 2001). Neste contexto, pode-se afirmar que a disciplina de Engenharia do Conhecimento está envolvida de alguma maneira em todas as etapas do processo de GC, e possui um papel importante no norteamo das ações da gestão.

De acordo com as definições apresentadas relativas ao conceito de governo eletrônico, quando relacionado ao conceito de conhecimento apresentado anteriormente, observa-se que a possibilidade do governo eletrônico em subsidiar a tomada de decisões, com base em conhecimentos, pode ser considerada de grande potencial no sentido de agregar valor em vários âmbitos envolvidos nas organizações públicas.

3. METODOLOGIA

A metodologia utilizada para este trabalho partiu de uma pesquisa documental, onde foram analisadas diversas fontes de dados abertos que pudessem ser utilizadas para o objetivo fim deste trabalho. Optou-se pela disponibilidade dos dados em se utilizar as fontes de dados do Instituto Brasileiro de Geografia e Estatística (IBGE) denominado Fluxos Aéreos 2010 (IBGE, 2015) por se tratarem de dados abertos de uma instituição federal brasileira, cujos dados tem ampla representação e abrangem todo o território nacional.

Selecionada a fonte dos dados optou-se pelo uso de softwares para tratamento destes dados, as tecnologias utilizadas fazem parte da suite de ferramentas da Microsoft, sendo o Sistema Gerenciador de Banco de Dados (SGBD), o SQL Server 2012 versão 11.0.5343, usado para criação da base de dados, e a ferramenta para acesso, configuração e manipulação desta base foi o SQL Server Management Studio versão 11.0.3128.0. Além disso também foram usadas a ferramenta Assistente de Importação e Exportação do SQL Server para extração dos dados de planilhas Excel e para o armazenamento na base de dados de destino.

No capítulo seguinte é apresentado o caso de aplicação prática do processo ETL utilizando os dados abertos, coletados do IBGE.

4. APLICAÇÃO DO PROCESSO ETL COM OS DADOS ABERTOS DO IBGE REFERENTES A MOVIMENTAÇÃO DE PASSAGEIROS E CARGAS AÉREAS DA AVIAÇÃO CIVIL BRASILEIRA

Os dados utilizados para aplicação do processo ETL são dados obtidos a partir de uma pesquisa denominada Ligações aéreas, realizada pelo IBGE, inserida dentro do projeto de Redes e Fluxo do instituto. Esta pesquisa abrange a movimentação de passageiros e cargas aéreas dos aeroportos brasileiros, sendo no total 138 municípios inseridos no escopo da última pesquisa realizada, no ano de 2010.

4.1. Extração dos dados

O processo de extração dos dados se iniciou através da busca da fonte primária, com base no domínio previamente estabelecido. Desta maneira, foram encontrados os dados disponibilizados na página web do IBGE, em formato XLS e ODB, para download, conforme Figura 1.

The screenshot shows the IBGE website interface. At the top, there are navigation links for 'BRASIL', 'Acesso à informação', 'Participe', and 'Serviço'. Below this is the IBGE logo and the motto 'ORDEM E PROGRESSO'. A search bar is visible on the right. The main navigation menu includes 'Indicadores', 'População', 'Economia', 'Geociências', 'Canais', 'Download', and 'Participação'. The current page is titled 'Geografia' and contains a sidebar with a tree view of categories such as 'Introdução', 'Divisão Regional', 'Divisão Urbano Regional', 'Regiões Rurais', 'Evolução da Divisão Territorial do Brasil 1872-2010', 'Banco de Dados', 'Áreas Especiais', 'Amazônia Legal', 'Faixa de Fronteira', 'Semiárido', 'Zona Costeira', 'Atlas', 'Mapas', 'Regiões de Influência das Cidades', 'Publicação', 'Base de Dados', 'Geografia Urbana', 'Áreas Urbanizadas', 'Mapas (formato PDF)', 'Mapas (formato SHP)', 'Arranjos Popacionais e Concentrações Urbanas do Brasil', 'Publicação completa (formato PDF)', 'Tabelas', 'Base de Dados', 'Mapas (formato PDF)', 'Mapas interativos', 'Redes e Fluxos do Território', 'Ligações Aéreas 2010', 'Publicação completa (em formato PDF)', and 'Base de Dados'. The main content area is titled 'Ligações Aéreas - 2010' and includes a 'Tabelas' section with the following text: 'Os dados levantados e analisados no projeto Redes e Fluxos no Território - As Ligações Aéreas 2010 estão disponibilizados em planilhas, descritas a seguir:'. Below this text is a list of data sources: 'Banco de dados das cidades', 'Ligações aéreas-passageiros e Ligações aéreas-carga', 'Fluxos principais', 'Modelo gravitacional', and 'Matriz de custo e Matriz de tempo'. A table at the bottom of the page lists the available files in XLS and ODS formats.

Arquivos em formato .XLS	Arquivos em formato .ODS
Banco de dados das cidades	Banco de dados das cidades
Fluxos principais	Fluxos principais
Ligações aéreas_carga	Ligações aéreas_carga
Ligações aéreas_passageiros	Ligações aéreas_passageiros
Matriz de custo	Matriz de custo
Matriz de tempo	Matriz de tempo
Modelo gravitacional	Modelo gravitacional

Figura 1: Página do Instituto Brasileiro de Geografia e Estatística (IBGE) Fonte: IBGE, 2015, disponível em:

http://www.ibge.gov.br/home/geociencias/geografia/redes_fluxos/ligacoes_aereas_2010/base.shtm

Foi constatada a existência de sete arquivos disponíveis para download, sendo realizada uma análise preliminar no conteúdo destes arquivos, dado o escopo deste trabalho, conclui-se que os arquivos a serem utilizados são o Banco de dados das cidades, e os arquivos de Ligações aéreas_carga e Ligações aéreas_passageiros. Sendo o primeiro arquivo utilizado para montar o banco de dados com informações de municípios, e os dois últimos para obter os registros de movimentações, por ligações aéreas, ou seja por pares de origem e destino.

Para os demais arquivos disponíveis, identificou-se que os dados não são relevantes nesta pesquisa, pois não contém dados de movimentação.

Entre os formatos disponíveis para download dos arquivos escolhidos, optou-se pelo uso do formato XLS, porém, cabe salientar, que no caso do arquivo dos dados de cargas aéreas movimentadas, foi verificado que o arquivo em formato XLS não estava disponível, sendo necessário a utilização do arquivo no formato ODB.

Com os dados em posse dos pesquisadores, se iniciou o processo de transformação, onde estes dados foram analisados detalhadamente, e preparados para uma utilização de forma estruturada.

4.2. Transformação dos dados

A transformação de dados se iniciou a partir de uma análise criteriosa nas colunas disponibilizadas em cada arquivo, sendo identificado no arquivo de cidades.xls a existência de duas planilhas, onde a primeira é chamada de “Banco de dados”, da qual foram encontradas as seguintes colunas:

- UF
- CodUF
- Codmundv
- Codmun
- nomemun
- População residente
- Produto interno bruto municipal
- Níveis de centralidade 2007
- Denominação de centralidade 2007
- Quantidade de passageiros movimentados em 1972
- Quantidade de carga movimentada em 1972
- Quantidade de passageiros movimentados em 1975
- Quantidade de carga movimentada em 1975
- Quantidade de passageiros movimentados em 1980
- Quantidade de carga movimentada em 1980
- Quantidade de passageiros movimentados em 1985
- Quantidade de carga movimentada em 1985
- Quantidade de passageiros movimentados em 1990
- Quantidade de carga movimentada em 1990
- Quantidade de passageiros movimentados em 1995
- Quantidade de carga movimentada em 1995
- Quantidade de passageiros movimentados em 2000
- Quantidade de carga movimentada em 2000
- Quantidade de passageiros movimentados em 2005
- Quantidade de carga movimentada em 2005
- Quantidade de passageiros movimentados em 2010
- Quantidade de carga movimentada em 2010
- Custo médio das viagens a partir da origem
- Tempo médio das viagens a partir da origem

E na segunda planilha nomeada como “Dicionário de variáveis”, foram encontradas as seguintes colunas:

- Código/Sigla
- Descrição
- Fonte
- Ano
- Unidade de medida

Com esta análise foi possível selecionar quais as colunas relevantes para este trabalho, verificando na primeira planilha citada a existência de dados pertencentes exclusivamente a municípios, tais como os dados de: UF, CodUF, Codmundv, Codmun, nomemun, População residente, Produto interno bruto municipal, Níveis de centralidade 2007 e Denominação de centralidade 2007. E na segunda planilha, identificou-se os dados referentes as fontes primárias, servindo como um metadado da planilha anterior, neste caso todas as colunas foram selecionadas. Cabe salientar, que os dados de movimentação de passageiros e cargas da primeira planilha não foram utilizados, uma vez que estas informações constam de forma detalhada por ligação aérea nos outros arquivos baixados.

Quanto aos outros dois arquivos baixados, sendo respectivamente referentes aos dados de movimentação de passageiros e de cargas, ambos agrupados por ligações aéreas (município de origem e destino), foi verificado a existência de diversas planilhas com a mesma estrutura, porém separadas

pelos anos de coleta da pesquisa, sendo os anos de 1972, 1975, 1980, 1985, 1990, 1995, 2000, 2005 e 2010. Sobre a estrutura, constam as seguintes colunas:

- ID;
- Cod-origem;
- Mun-origem;
- Cod-destino;
- Mun-destino; e
- Pax - no arquivo de movimentação de passageiros e Carga(kg) - no arquivo de cargas movimentadas.

Com o conteúdo analisado e as colunas selecionadas, foi possível realizar a normalização dos dados a partir da elaboração de um modelo de entidade relacional, que possibilita o armazenamento consistente e um eficiente acesso aos dados, conforme apresentado na Figura 2.

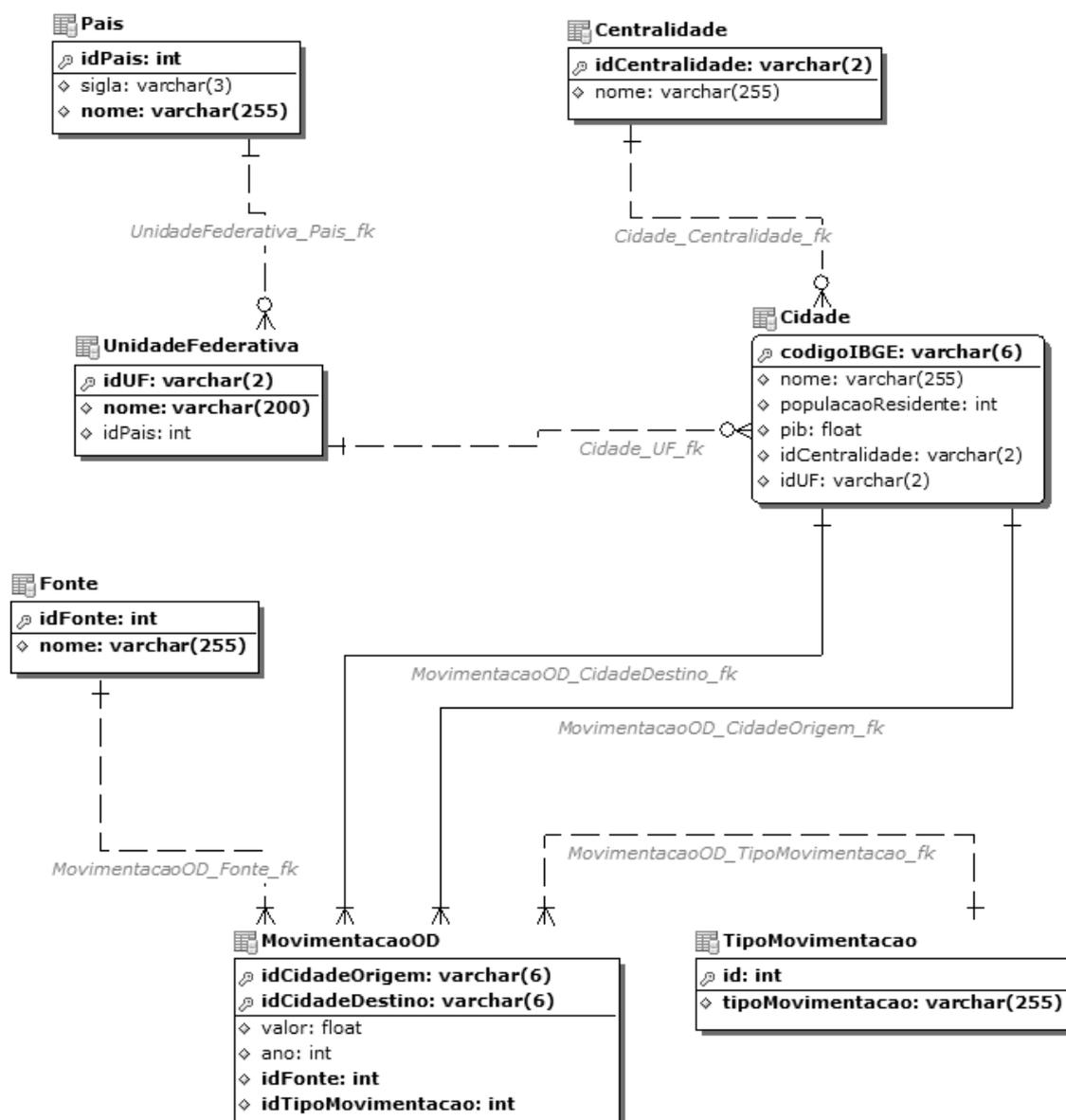


Figura 2: Modelo de Entidade Relacional Fonte: Elaborado pelos autores

Para facilitar a carga dos dados na base, foram gerados *templates* em Excel com uma estrutura similar ao modelo de dados criado, para que os dados possam ser transferidos do modelo disposto pelo IBGE, para uma estrutura mais próxima da do modelo elaborado. Desta maneira, o processo de carga dos dados pode ser otimizado, conforme é possível conferir no tópico a seguir.

Vale ressaltar uma inconsistência encontrada no arquivo de movimentação de passageiros, no ano de 1990, onde existem dois registros para um mesmo par de origem e destino, sendo a origem o município de São Paulo e destino Belo Horizonte. Com a modelagem normalizada estes registros não poderiam ser importados, daria erro, pois a chave da tabela MovimentacaoOD são os próprios códigos de municípios, com mais as informações de tipo de movimentação e ano. Desta maneira, optou-se pela soma dos dois registros de movimentação, sendo inserido no *template* para importação um único registro para este caso, respeitando a consistência da base de dados.

4.3. Carga dos dados em base de dados relacional

O Sistema de Gerenciamento do Banco de dados (SGBD) utilizado foi a ferramenta SQL Server 2012 versão 11.0.5343. Neste servidor foi criada uma base de dados cuja a estrutura de tabelas e relacionamentos estão de acordo com o modelo elaborado, previamente apresentado na Figura 2.

Para a realização da carga dos dados efetivamente na base de dados criada, foi utilizado o Assistente de Importação e Exportação da plataforma SQL Server, o qual possibilita a configuração da origem e destino do dado, que neste caso teve como origem os *templates* excel preenchidos com os dados transformados do IBGE, e como a fonte de destino à base de dados criada para realização deste trabalho. Neste momento, é necessário o apontamento detalhado a nível de colunas e tabelas, para que o dado possa ser corretamente carregado no novo modelo de dados.

Vale ressaltar que este processo de importação deve respeitar as hierarquias e consistências expostas no modelo de dados, sendo necessária rodar várias vezes o assistente para importar parte a parte os dados, respeitando todos os relacionamentos. Por exemplo, não é possível a importação dos dados de cidade, sem ter os registros na tabela de UF.

Por fim, uma quantificação dos registros consolidados na base de dados, pode ser visualizada na Tabela 1.

Tabela 1: Número de registros importados no processo de ETL. Elaborado pelos autores

Tabela	Quantidade de registros importados
País	1
UF	28
Cidade	138
Centralidade	11
Fonte	5
TipoMovimentacao	2
MovimentacaoOD	2046

Com uma base de dados consolidada e estruturada, torna-se mais fácil a realização de agrupamentos, cruzamento de dados para geração de mapas e gráficos ou qualquer outro tipo de representação do conhecimento, uma vez que as consultas aos dados podem ser feitas de forma customizáveis e rápidas, o fato de utilizar a linguagem SQL¹ nas consultas dá mais rapidez.

¹ Structured Query Language: linguagem padrão para pesquisa em bases de dados relacionais.

4.4. Representação e descoberta de conhecimentos

A partir dos dados estruturados, foi possível a realização de algumas consultas nesta base, buscando demonstrar o potencial em gerar diferentes visões e representá-las através de gráficos, possibilitando a descoberta de novos conhecimentos.

No gráfico exposto na Figura 3, são apresentados os totais de movimentação de passageiros registrados em cada pesquisa realizada, ao longo dos anos.

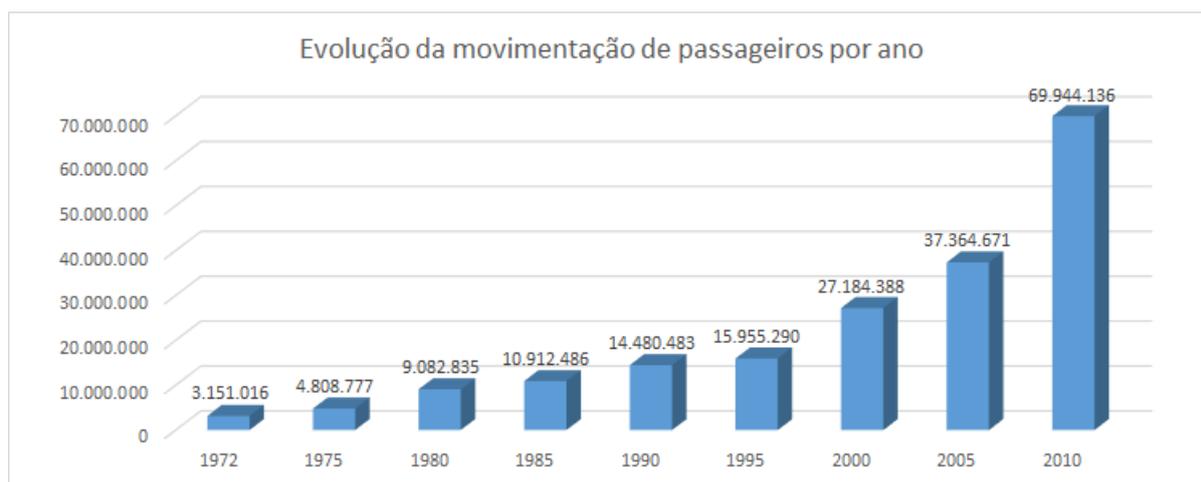


Figura 3: Gráfico de evolução da movimentação de passageiros por ano. Elaborado pelos autores.

Outra representação foi gerada no intuito de apresentar o número de ligações aéreas entre municípios, que estão inseridos em cada ano em que as pesquisas foram realizadas, sendo apresentado o resultado no gráfico da Figura 4.

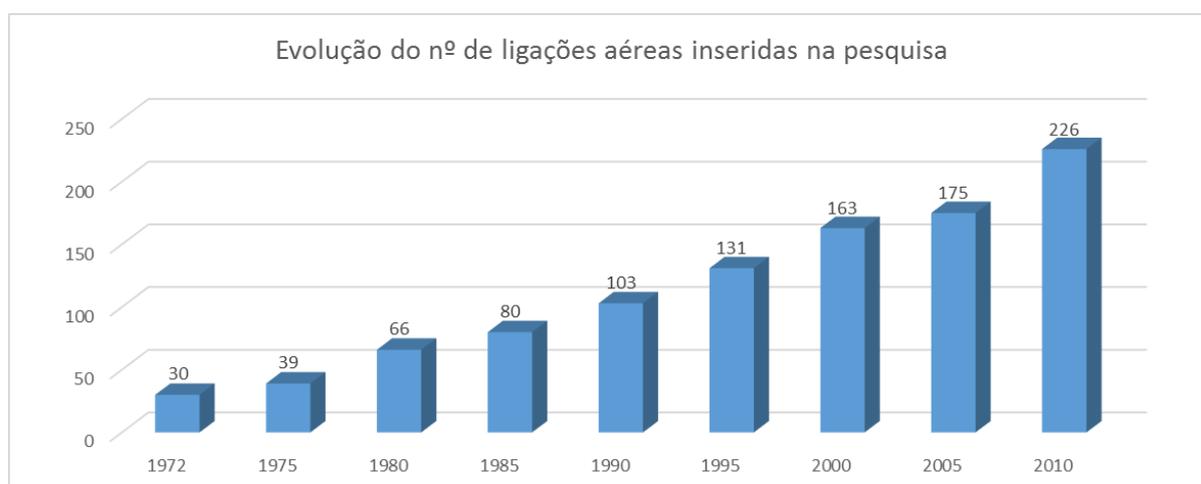


Figura 4: Gráfico de evolução do número de ligações aéreas nas pesquisas realizadas. Elaborado pelos autores.

Visando evidenciar os dados mais atuais da pesquisa do IBGE, sendo referentes ao levantamento realizado no ano de 2010, foi elaborado um gráfico de movimentação geral de passageiros, onde estão apresentados todos os municípios com suas respectivas movimentações,

conforme Figura 5. Neste gráfico é possível identificar, por exemplo, que as três cidades mais relevantes tratando-se de movimentação, são respectivamente: São Paulo, Rio de Janeiro e Brasília.

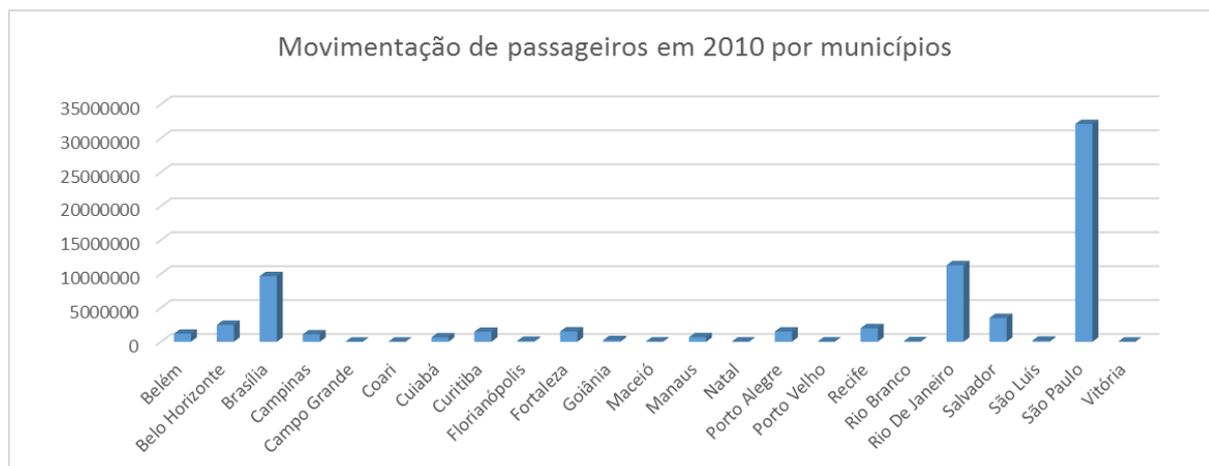


Figura 5: Movimentação de passageiros por municípios de origem em 2010. Elaborado pelos autores.

As representações de conhecimento elaboradas são apenas exemplos que servem para demonstrar o grande potencial em trabalhar com os dados abertos do governo, e com um processo estruturado, utilizando técnicas e ferramentas consolidadas da área de tecnologia da informação.

5. RESULTADOS

Com a execução deste trabalho, foram gerados diversos resultados, como uma base de dados consolidada com todos os dados desta pesquisa do IBGE, possibilitando inclusive que ao ser realizada uma atualização destes dados seja possível incrementar os mesmos nesta base, e ampliar ainda mais a gama de registros disponíveis no banco de dados.

Além disso, como resultado também foi gerado um modelo de entidade-relacional que é conceitual, e possibilita que o mesmo tenha sua estrutura criada em qualquer outro SGBD, ou até mesmo adaptado para abranger novos conceitos e ampliar ainda mais o modelo proposto e validado neste trabalho.

Outro resultado interessante foi o levantamento, utilização e comprovação de uma série de ferramentas e técnicas que auxiliam na realização do processo de extração, transformação e carga dos dados (ETL).

6. CONCLUSÕES

A cada vez mais o governo está tomando iniciativas em disponibilizar os dados abertos para a população. Isso permite maior transparência nas ações do governo, bem como possibilita a descoberta de novos conhecimentos através de pesquisas científicas com estes dados.

Com a possibilidade de criar diferentes visões, cruzamento de dados e novos cenários para análise, através de consultas automatizadas nos dados consolidados do IBGE, se torna ágil e viável a geração de novos resultados, por exemplo. Pois este trabalho sendo realizado diretamente em planilhas do Excel, sem a utilização de uma base de dados relacional, poderia inclusive gerar problemas futuros, devido a limitações em número de registros aceitos pelo Excel em relação a grande massa de dados geradas nestas pesquisas públicas. Neste caso, conclui-se que é possível a expansão do modelo conceitual criado, para abranger também o escopo de outras pesquisas do governo, aumentando ainda mais as possibilidades de descoberta de novos conhecimentos.

A execução deste trabalho também possibilitou concluir que a utilização de ferramentas e técnicas adequadas para execução de um processo de ETL, podem auxiliar de forma eficiente no processo propriamente dito.

7. REFERÊNCIA BIBLIOGRÁFICA

Abreu, F.S.G.G. (2008). Desmistificando o conceito de ETL. *Revista de Sistemas de Informação*, 2.

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, 107-136.

Eletrônico, G. (2015). Padrões de Interoperabilidade de Governo Eletrônico–ePing.

Grant, G. (2005) Realizing the promise of electronic government. *Journal of Global Information Management*, v. 13, n. 1.

IBGE, 2015, disponível em: http://www.ibge.gov.br/home/geociencias/geografia/redes_fluxos/ligacoes_aereas_2010/base.shtm

Kimball, Ralph. *Data Warehouse Toolkit*. Tradução Mônica Rosemberg; Revisão Técnica Ronal Stevis Cassiolato. São Paulo: Makron Books, 1998.

Koh, C. E., & Prybutok, V. R. (2003). The three ring model development of an instrument for measuring dimensions of e-government functions. *The Journal of Computer Information Systems*, 43(3), 34.

Rautenberg, S. (2009). Modelo de conhecimento para mapeamento de instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias.

Rotta, M.J.R. (2013). Modelagem do conhecimento legal necessário para elaboração de sentenças em processo na área de defesa do consumidor.

Rover, A. J. (2008). O governo eletrônico e a inclusão digital: duas faces da mesma moeda chamada democracia. *Inclusão digital e governo eletrônico*. Zaragoza: Prensas Universitárias de Zaragoza, Lefis series, 3, 9-34.