Faculdade de Engenharia da Universidade do Porto

Extracção de Conhecimento no Observatório Social de Santa Maria da Feira

Luís Filipe Ferreira de Bastos

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major Telecomunicações

Orientador: Prof. Dr. Francisco José de Oliveira Restivo

Julho de 2010

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES 2009/2010

A Dissertação intitulada

"EXTRACÇÃO DE CONHECIMENTO NO OBSERVATÓRIO SOCIAL DE SANTA MARIA DA FEIRA"

foi aprovada em provas realizadas em 22/Julho/2010

o júri

Presidente Professora Doutora Ana Maria Rodrigues de Sousa Faria de Mendonça
Professora Associada do Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professor Doutor Paulo Sérgio Tenreiro de Magalhães
Professor Auxiliar da Universidade Católica Portuguesa

Professor Doutor Francisco José de Oliveira Restivo
Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto (Orientador)

O autor declara que a presente dissertação (ou relatório de projecto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extractos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são correctamente citados.

Autor - **Luís Filipe Ferreira de Bastos**

Faculdade de Engenharia da Universidade do Porto

Resumo

Com o passar dos anos, a capacidade de processar, analisar e explorar informação tem vindo a ser ultrapassada pela capacidade de a recolher e armazenar. Muitas empresas, associações ou organizações, têm agora a prática de recolher grandes quantidades de informação de um modo contínuo, seja sobre as suas operações, produtos, clientes, pesquisas, trabalhos, etc., sendo a sua análise e processamento essenciais para tomadas de decisão.

Os observatórios sociais não fogem à regra. Possuindo grandes quantidades de informação sobre os mais variados aspectos sociais, criam vários problemas na organização, na estrutura, e principalmente na visualização dos seus dados, dada a sua natureza e o tipo de utilizador. Este documento retrata alguns destes problemas no Observatório Social de Santa Maria da Feira.

Idealizado pela Divisão Social da Câmara Municipal de Santa Maria da Feira, o Observatório abarca cinco áreas: Educação, Emprego, Família, População e Comportamentos Desviantes, e inclui centenas de indicadores. Possui como objectivo apresentar a sua informação de forma a facilitar ao máximo a sua visualização e a descoberta de conhecimento por parte de utilizadores com formação específica na área social mas sem desprezar o utilizador comum.

O Observatório é um sistema baseado numa base de dados computadorizada que guarda, processa e manipula informação espacial referente ao concelho de Santa Maria da Feira. Esta informação possui também uma componente temporal, uma vez que varia ao longo do tempo - anos, décadas, ou qualquer outra granularidade temporal.

Ao longo desta dissertação são analisados os vários problemas que o Observatório Social suscita, dando-se ênfase à visualização de dados. Neste contexto, um dos principais objectivos é o desenvolvimento de uma interface para a visualização dos dados dos indicadores que permita representar dados espaço-temporais com uma, duas, três ou mesmo quatro dimensões, cuja possibilidade foi implementada no decorrer desta dissertação.

Esta tese aborda também a extracção de conhecimentos do Observatório, globalmente denominado por *Web Mining,* nomeadamente padrões de utilização, de forma a enriquecer o conhecimento das entidades que intervêm no terreno social. Serve igualmente para melhorar o Observatório, estruturalmente e em conteúdo.

Neste documento são então apresentados os vários problemas a e as soluções implementadas. São também apresentadas as melhorias efectuadas, como é o caso da implementação da gestão de utilizadores, que o Observatório não possuía.

Abstract

Over the years, the ability to process, analyze and exploit information has been superseded by the ability to gather and store. Many companies, associations or organizations have now the practice of collecting in a continuous way, large amounts of information, either on its operations, products, customers, researches, works or others, being the analysis and processing essential for decision making.

The social observatories are no exception. Having large amounts of data on various social aspects, sometimes there are many problems in their organization, structure and mainly in their data visualization given their nature and the type of users. This document reflects some of these problems in the Social Observatory of Santa Maria da Feira.

Conceived by the Social Division of Santa Maria da Feira City Council, the Observatory encompasses five areas: Education, Employment, Family, Population and Deviant Behavior, and includes hundreds of indicators. Its main objective is to present information in a friendly and clean way to allow straightforward visualization and knowledge extraction, in order to facilitate users with specific training social areas, without neglecting regular users.

The Observatory is a computerized system based on a Database that stores, processes and manipulates spatial information relating to the Santa Maria da Feira municipality. This information also has a temporal component, since it varies over time – years, decades, or any other temporal granularity.

Throughout this thesis it´s described the range of problems that the social observatory raises, with emphasis on data visualization. In this context, one of the main objectives it´s the development of an interface for viewing the indicators data, that allows space-time data, associated to one, two, three or even four dimensions, which possibility was implemented during this thesis.

This thesis also refers to the Observatory knowledge extraction, generally referred as Web Mining, namely usage patterns, in order to enrich the entities involved in social field knowledge. It also serves to improve Observatory structure and content.

So, this document presents the various problems analyzed and the solutions implemented. It also presents the improvements, such as the implementation of system user management which the observatory had not.

Agradecimentos

Em primeiro lugar gostaria de agradecer às pessoas que mais me apoiaram ao longo do curso de Engenharia Electrotécnica e de Computadores.

Agradeço em tom especial aos meus Pais pelo apoio em todos os momentos, principalmente nos mais difíceis.

Agradeço especialmente ao meu Irmão por todo o auxílio prestado no decorrer do curso através dos seus megas conhecimentos.

Não posso deixar de agradecer à Margarida pelo apoio e carinho proporcionados.

Gostaria também de agradecer às pessoas que me ajudaram nesta dissertação em especial ao Professor Doutor Francisco José de Oliveira Restivo.

Agradeço igualmente às colaboradoras do Observatório Social de Santa Maria da Feira pelo auxílio prestado na validação dos resultados finais.

Um obrigado especial também à Raquel pela grande ajuda na revisão.

Por fim gostaria de agradecer a todos os amigos e colegas que me acompanharam durante estes anos nos momentos de estudo, de entreajuda, e nos momentos divertidos.

Índice

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

Lista de abreviaturas (ordenadas por ordem alfabética)

AJAX	Asynchronous JavaScript And XML
API	Application Programming Interface
SIG	Sistema de Informação Geográfica
GPL	General Public License
HTML	HyperText Markup Language
IDIT	Instituto de Desenvolvimento e Inovação Tecnológica
IP	Internet Protocol
PHP	Hypertext Preprocessor
SQL	Structured Query Language
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

Capítulo 1

Introdução

O Observatório Social de Santa Maria da Feira [38] foi desenvolvido pelo IDIT, Instituto de Desenvolvimento e Inovação Tecnológica, unidade à qual a Universidade do Porto está associada, e tem como objectivo disponibilizar a todos os interessados os dados que caracterizam socialmente o município de Santa Maria da Feira e as suas freguesias, onde centenas de entidades públicas e privadas realizam actividades de índole social.

Idealizado pela Divisão Social da Câmara Municipal de Santa Maria da Feira, abarca cinco áreas: Educação, Emprego, Família, População e Comportamentos Desviantes, e incluí centenas de indicadores.

O Observatório é um sistema baseado numa base de dados e uma interface Web que guarda, processa e manipula informação espacial referente ao concelho de Santa Maria da Feira. Esta informação possui também uma componente temporal, uma vez que varia ao longo do tempo - anos, décadas, ou qualquer outra granularidade temporal. Tem provado ser um instrumento valioso para as entidades que intervêm no terreno social, tendo criado oportunidades para trabalhos muito interessantes, que podem ir desde a sua revisão técnica e expansão a novas áreas temáticas, o estudo do impacto a nível da visualização da introdução de mais anos de observação, até ao desenvolvimento de novas técnicas de visualização de dados espaço-temporais, *data mining*, aprendizagem e filtragem colaborativas, etc.

O Observatório permite então uma análise, gestão e representação do espaço e dos fenómenos que nele ocorrem. É extremamente importante que essa análise, gestão e representação sejam feitas do melhor modo, de forma a possibilitar ao máximo a extracção de conhecimento por parte do utilizador comum e do utilizador experiente, como por exemplo, sociólogos.

Esta dissertação surge da necessidade de dar resposta a um grande problema que afecta a maioria dos observatórios sociais: a representação e visualização dos dados. Sendo extremamente complicado encontrar boas soluções para representar dados complexos de um modo simples, é necessário criar ou adaptar uma solução da melhor maneira possível para o ambiente de um observatório. Esta tese considera também que é necessário explorar a extracção de conhecimento no Observatório, o denominado *Web Mining,* através de padrões de utilização do sistema, de forma a enriquecer o conhecimento das entidades que intervêm no terreno social e também de forma a enriquecer o Observatório em conteúdo e estrutura.

1.1 - Descrição do Problema

Existem vários problemas que irão ser abordados nesta dissertação. Inicialmente é abordado o problema da visualização dos dados dos indicadores e consequentemente da estrutura dos indicadores. Seguidamente aborda-se o problema da extracção de conhecimento através dos padrões de utilização do Observatório. Existe também um problema secundário que se pretende resolver, nomeadamente a gestão dos utilizadores. Durante a dissertação não se excluem as resoluções de outros problemas inerentes ao seu desenvolvimento.

1.1.1 - Visualização

O Observatório Social é constituído por variados indicadores, estando estes divididos por áreas, grupos e temas. Estes indicadores apresentam dados espaço-temporais de uma forma que possibilita a sua utilização pelas pessoas, quer estejam envolvidas no meio social ou não, servindo como instrumentos para avaliação e comparação de forma a encontrar padrões e tendências. No Observatório a visualização dos dados é então extremamente importante, pois pretende-se que as pessoas que o visitam possam tirar o maior partido desses dados. O Observatório é capaz de representar dados com uma, duas ou três dimensões, sendo essas representações em forma de tabelas e gráficos, tal como está ilustrado no exemplo da figura 1.1.

	Ano lectivo 2004/2005		Ano lectivo 2005/2006	
	Nº alunos	Nº retenções	Nº alunos	Nº retenções
AV Argoncilhe	631	58	597	21
AV Arrifana	326	46	408	24
AV Canedo	393	16	388	17
AV Fernando Pessoa	1081		1045	32
AV Fiães	551	40	574	33
AV Lobão	604	28	580	10
AV Lourosa		16		
AV Paços Brandão	230	58	213	9
AV Ferreira Almeida	737	20	663	34
AV Milheirós Poiares	499	27	472	32
AH Nogueira, Mozelos e Lamas	863		870	37

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

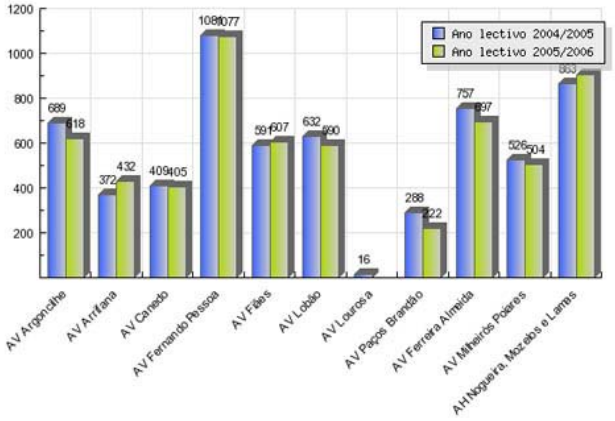


Figura 1.1: Exemplo de visualização de um indicador de três dimensões

Uma grande falha destes tipos de representações inicialmente presentes no Observatório é a sua abrangência limitada. No exemplo da figura 1.1 apenas estão presentes dois anos lectivos. Ao introduzir mais dados, as visualizações - quer nas tabelas, quer nos gráficos – tornam-se extremamente confusas, gerando problemas de interpretação. Na realidade, a visualização gráfica é incapaz de representar indicadores com três dimensões.

Outro dos problemas da visualização de dados espaço-temporais ocorre nos indicadores que possuem a dimensão "freguesias". A visualização aqui é feita através de mapas, como o da figura 1.2, e tal como no caso anterior, os administradores são obrigados a criar vários indicadores para as diferentes dimensões temporais. Isto acontece por não haver forma de demonstrar no mesmo indicador os diferentes mapas para os diferentes registos temporais.

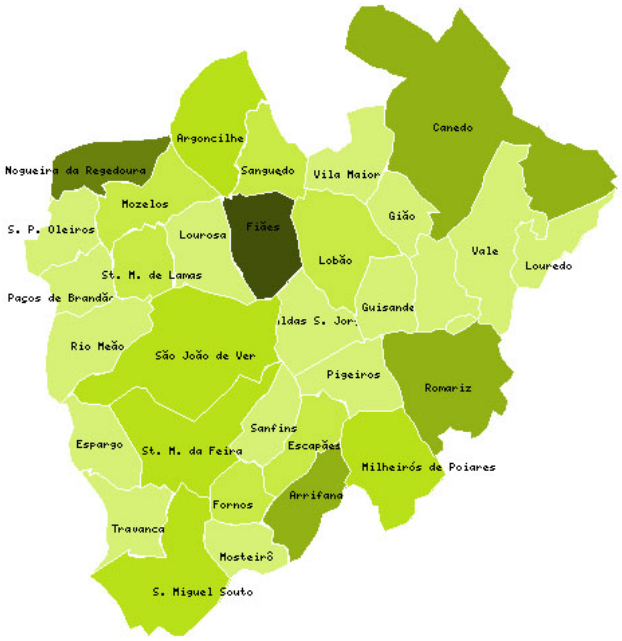


Figura 1.2: Exemplo de visualização em mapas

1.1.2 - Padrões de Utilização

Estando o Observatório Social em constante evolução e crescimento, é do interesse das pessoas nele envolvidas, oferecer aos utilizadores os conteúdos que mais lhes possam interessar. Por outro lado como se trata de um ambiente social, pode ser importante o enfoque às várias situações sociais, através da identificação das áreas ou indicadores que mais preocupam os cidadãos. Qualquer que seja a finalidade é necessário implementar uma plataforma que permita extrair informação relevante sobre os utilizadores e os seus hábitos, nomeadamente os percursos entre indicadores.

1.1.3 - Gestão de Utilizadores do Painel de Administração

Na sua versão inicial, no BackOffice do Observatório Social de Santa Maria da Feira, existe apenas um tipo de utilizador, o administrador. Este, é o responsável por toda a gestão do Observatório, tendo que criar, alterar ou apagar áreas, temas, fontes, dimensões, grupos e indicadores. É também o responsável pelos dados inseridos nos indicadores. O problema de

haver apenas um tipo de utilizador responsável pela gestão do Observatório, é que nem sempre as pessoas que inserem dados estão qualificadas para criar indicadores, pois estes seguem um determinado tipo de regras que não são de conhecimento geral. Novos utilizadores podem vir a alargar a interactividade do Observatório.

1.2 - Objectivos

A presente dissertação tem como um dos objectivos principais, o desenvolvimento de uma solução para a visualização dos dados dos indicadores. Sendo estes dados espaço-temporais, há que ter em conta a sua granularidade específica, criando soluções adequadas e próprias para os utilizadores alvo, que serão utilizadores com formação específica na área social. No entanto, os utilizadores comuns sem qualquer formação nesta área não deverão ser desprezados.

A interface terá que possuir tabelas com uma representação gráfica que suporte até quatro dimensões. Antes do desenvolvimento dessa solução, será necessária a implementação da construção de indicadores com quatro dimensões. Irá também ser necessário resolver alguns problemas na estrutura do Observatório. Pretende-se arranjar uma solução para o problema dos indicadores que possuam a sua visualização em mapas, de forma a ser possível observar no mesmo indicador os diferentes registos temporais e preencher algumas lacunas que este modo de visualização apresenta.

Outro dos objectivos será a implementação de uma solução para a extracção e registo de dados da navegação dos utilizadores para posterior apresentação. Deverá ser estudado o melhor modo de a fazer. Pretende-se também estudar a implementação da gestão de utilizadores no Observatório e os seus campos de aplicação.

Em complemento tentar-se-á também introduzir algumas técnicas de aprendizagem colaborativa. No caso de se revelar necessário e profícuo, não se excluem alguns aperfeiçoamentos da solução existente.

1.3 - Organização do Documento

Esta dissertação é constituída por cinco capítulos.

O primeiro capítulo, em que este texto se insere, é o capítulo introdutório onde se apresenta o contexto da dissertação assim como a descrição do problema e dos objectivos propostos.

O segundo capítulo resulta do trabalho de investigação efectuado acerca dos assuntos relativos à dissertação, nomeadamente a visualização, os dados espaços temporais e padrões de utilização, apresentando o actual estado de arte.

No terceiro capítulo é feita uma análise detalhada do Observatório Social de Santa Maria da Feira, dos problemas que este possui, dos objectivos a cumprir e os passos efectuados para a resolução destes.

O quarto capítulo apresenta a implementação efectuada para cumprir os vários objectivos propostos, apresentando os seus respectivos resultados.

No quinto e último capítulo é efectuada a recolha das principais conclusões retiradas dos capítulos anteriores e são propostos temas de trabalho que possam dar continuidade ao estudo aqui desenvolvido.

1.4 - Dados usados na Dissertação

Para a execução desta dissertação utilizaram-se dados puramente fictícios. Apesar de em alguns casos, os dados apresentados nas figuras serem os mesmos que se encontram no Observatório Social de Santa Maria da Feira, devem ser considerados como fictícios pois apenas foram colocados em ambiente de testes não sendo possível ao leitor distinguir quais os dados que serão verdadeiros. Em alguns casos os dados nem se encontram disponíveis.

Capítulo 2

Os Dados Espaço Temporais, a Visualização e os Padrões de Utilização

Este capítulo consiste na pesquisa e documentação do estado actual de desenvolvimento dos sistemas e subsistemas relacionados com o tema da dissertação. Como tal, são abordados vários assuntos relacionados. Inicialmente é feito um estudo de como funcionam os sistemas que possuem dados espaço-temporais e consequentemente as bases de dados espaço-temporais. De seguida é focada a visualização neste tipo de sistemas mostrando alguns exemplos de aplicação. Por fim é feito uma revisão geral do estado actual no campo da extracção de padrões de utilização.

2.1 - Introdução

Desde a década de 80, os desenvolvimentos em torno dos dados geoespaciais digitais têm ganho ímpeto, o que faz com que cada vez mais os utilizadores observem este tipo de dados como fonte de informação e conhecimento privilegiado. Com o computador surgiram as bases de dados e a possibilidade de gerar mapas, através da extracção de dados via *queries,* que permitiam fazer análises mais completas e aprofundadas do que com os tradicionais mapas físicos. Os programas que guardavam os dados geoespaciais em bases de dados e que permitiam a sua análise foram denominados de Sistemas de Informação Geográficos (*Geographic Information Systems)*.

O Observatório Social de Santa Maria da Feira é um sistema que pretende associar a informação proveniente de um SIG, referente a este município e às suas freguesias, ou seja, informação espacial, com dados obtidos no terreno relativos a um certo número de temas sociais relevantes, permitindo uma análise, gestão e representação do espaço e dos fenómenos que nele ocorrem.

O concelho é constituído por 31 freguesias, tendo o Observatório disponível informação disposta por freguesia e concelho, figura 2.1.

Figura 2.1: Mapa do concelho de Santa Maria da Feira

2.2 - Sistema de Informação Geográfica

Na sua definição mais simples, um SIG é um sistema de informação processado por computador que guarda, processa, e manipula informação espacial permitindo uma análise, gestão e representação do espaço e dos fenómenos que nele ocorrem.

O centro de um SIG é a base de dados, como se pode observar na figura 2.2. Um elemento importante destas bases de dados é o seu modelo. Numa aplicação normal, as bases de dados tem um modelo relativamente simples, mas na maioria dos SIG tal não acontece. Em [1] o autor explica o quão complexo estes modelos podem ser e também como estes podem ser representados em sistemas de informação.

Para um armazenamento e recuperação de dados eficiente, os dados necessitam de estar bem estruturados e representados, assim como os algoritmos que os operam.

Uma característica importante de um SIG é a capacidade de partilhar dados com diferentes centros de informação ou entre componentes, sendo estes dos mais variados dentro do mesmo sistema de informação.

Os dados não serão muito úteis se não forem transmitidos às pessoas que deles necessitam. Enquanto nos SIG normais os dados são apresentados apenas em tabelas ou comentários, nos SIG com dados geoespaciais, existe um maior leque de tipos de representação, como representações em mapas, ou outras formas mais dinâmicas. Apesar de todas as demais funcionalidades que um SIG possa apresentar, é necessário ter igualmente presente as suas limitações, de forma a conseguir alcançar o máximo proveito.

Os programas SIG necessitam de ter um nível elevado de funcionalidades de modo a satisfazer o grau elevado de análises e decisões.

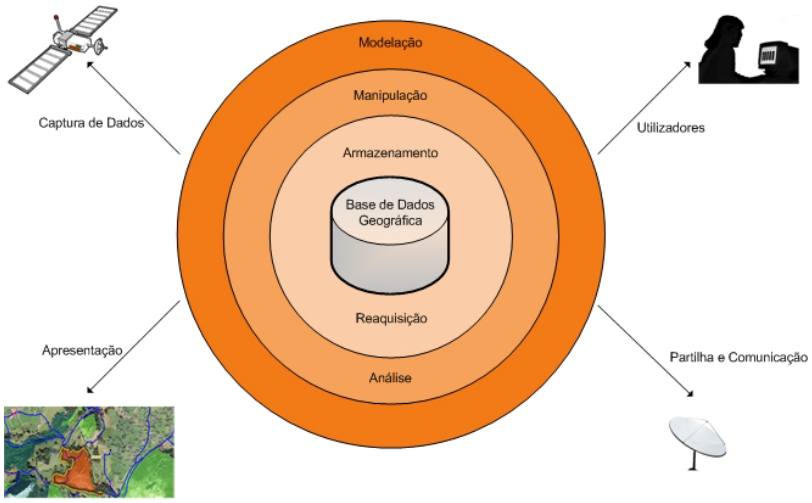


Figura 2.2: Constituição de um SIG

2.3 - Bases de Dados Espaço-Temporais

Com o passar dos anos, o avanço nas tecnologias de bases de dados e de gráficos tem vindo a possibilitar o desenvolvimento de sistemas capazes de lidar com dados relacionados com o espaço. Tal avanço fornece um vasto leque de aplicações, tais como: a monitorização de locais específicos, como florestas e oceanos; o alojamento de dados para caracterização estatística ou social de determinados locais, como é o caso deste projecto. Contudo, esta informação não é apenas espacial pois uma vez que o mundo não é estático, ela varia com o tempo. Estas duas componentes estão directamente relacionadas. Como referido anteriormente, os dados são recolhidos num determinado espaço, domínio espacial, e acontecem num determinado tempo, domínio temporal. Em grande parte dos sistemas esta dimensão temporal é representada através de momentos, de X em X tempo - como é o caso do Observatório, onde os dados são apresentados em vários registos temporais e não de um modo contínuo. Para os dados serem representados de um modo contínuo é necessário unir os dois domínios. Em [2], *Worboys*, fala do trabalho feito em modelação de informação que referencia uma única dimensão, a espaço-temporal.

Neste campo existem inúmeras aplicações que beneficiam de *queries* espaço-temporais, como por exemplo, sistemas de comunicações móveis, sistemas de controlo de tráfego, os SIG ou até aplicações multimédia.

Existem vários factores a ter em conta quando se considera a componente temporal para além da espacial:

- A localização de certo objecto pode mudar ao longo do tempo, por exemplo, se a função do sistema for monitorizar o tráfego automóvel;
- A geometria do objecto a ser medido pode mudar ao longo do tempo, por exemplo, a expansão de uma área/local;
- Os atributos do objecto podem mudar ao longo do tempo, por exemplo, o proprietário de certo objecto;

No caso concreto do Observatório apenas são consideradas as alterações dos atributos ao longo do tempo, alterações sócio económicas como o desemprego e o insucesso escolar, e não alterações na geometria ou localização.

Existem inúmeros casos onde este tipo de sistema pode ser aplicado, como por exemplo na análise da ocorrência de crimes [40], análise de tráfego automóvel [41], registo de acidentes de automóveis [42], análises geológicas e biológicas [43], entre muitos outros. Na sua maioria estes SIG requerem um estudo aprofundado da área de aplicação para uma melhor representação dos dados, que irá permitir encontrar respostas a questões levantadas ou encontrar soluções para os problemas em que se inserem.

2.4 - Dados Espaço-Temporais

Depois de uma introdução aos SIG espaço-temporais e da sua contextualização nesta tese, é necessário perceber como se trabalha com os dados que estes sistemas albergam. Tal como referido anteriormente, o Observatório contêm vários tipos de dados, que variam no tempo, referentes ao concelho de Santa Maria da Feira. Como o objectivo deste trabalho é o desenvolvimento de uma solução para a visualização destes dados, torna-se essencial entender como estes dados são guardados e a melhor forma de lidar com eles. *Andrienko* em [3] refere que *Peuquet* em [6] distingue este tipo de dados em três componentes. Espaço, *where*, Tempo, *when*, e Objectos, *what*:

- *when* + *where* -> *what*: Descreve o objecto ou conjunto de objectos que estão presentes numa dada localização ou conjunto de localizações num determinado tempo ou conjunto de tempos;
- *when* + *what* -> *where:* Descreve a localização ou conjunto de localizações ocupados por um dado objecto ou conjunto de objectos num determinado tempo ou conjunto de tempos;
- *where* + *what* -> *when*: Descreve o tempo ou conjunto de tempos em que um dado objecto ou conjunto de objectos ocupavam numa determinada localização ou conjunto de localizações.

Outras divisões deste tipo de dados foram feitas. Mesmo em [3], *Andrienko* dá a sua abordagem às diferentes categorias em que se podem dividir os dados espaço-temporais.

2.4.1 - Modelos

Existem vários modelos na literatura para os diferentes dados espaço-temporais. Os modelos espaço-temporais são o núcleo destes sistemas. Eles definem os tipos de dados dos objectos, as relações, as operações e as regras para manterem a integridade da base de dados. Estes modelos são concebidos com a intenção de lidarem com aplicações do mundo real, onde alterações espaciais ocorrem durante o tempo. Num modelo o tempo é representado por marcas temporais, sendo que cada modelo tem a sua representação distinta.

Os primeiros modelos a aparecer foram o *"Time Slice or Snapshot Model"* de *Armstrong* [4] e o *"Space-Time Composites"* de *Langran and Chrisman* [5], ambos em 1988. Em 1992 *Worboys* surgiu com o modelo *"Spatial Objects"* [2].

- *Time Slice or Snapshot Mode*l – Neste modelo são guardadas, na base de dados, camadas temporais independentes umas das outras. Todas as alterações que acontecem nos objectos espaciais são geridas com estas camadas, com uma granularidade predefinida. Por exemplo, numa região, os dados do desemprego

para o ano 2000 foram recolhidos e guardados na base de dados numa camada. Noutra camada são guardados os dados do desemprego para o ano 2001. O senão deste modelo é não ser totalmente eficiente a lidar com *queries* temporais.

- *Space-Time Composites* – Este modelo estende o de *Armstrong* ao sobrepor todas as camadas temporais numa única camada espaço-temporal. Cada formação espaço-temporal tem o seu único rumo de alterações nos atributos. Descreve as alterações de um objecto espacial durante um período de tempo. Uma formação é diferente de todas as formações adjacentes, o que faz com que a configuração do SIG necessite de ser reconstruída sempre que novos dados são introduzidos, visto que as formações vão ser alteradas devido às novas sobreposições.
- *Spacial Objects* – É uma extensão do anterior para uma dimensão 3D espaço-temporal de objectos. Não necessita de reconfigurações para a introdução de novos dados, contudo repete informação não alterada.

Há Autores que sugerem melhorias nestes modelos ou então modelos novos, como o *"Event Based Space Time Data Model"* por *Peuquet* e *Duan* [6] e o *"Three Domain Level"* por *Yuan* [7]. Mais tarde, tentando responder aos requisitos que foram despontando, vários Investigadores propuseram modelos para casos em que os objectos mudam de uma forma contínua, "*Moving Object Data Models*" [11]. Todos os modelos têm diferenças entre si, apresentando vantagens e desvantagens. Em [12] é feita uma análise aos modelos de bases de dados espaço-temporais existentes, comparando-os nos diferentes domínios.

2.5 - O Problema da Visualização

Com o passar dos anos, a capacidade do processar, analisar e explorar informação tem vindo a ser ultrapassada pela capacidade de recolher e armazenar informação. Muitas empresas e organizações têm agora o hábito de recolher rotineiramente grandes quantidades de informação com enfoque nas mais variadas áreas de actuação - operações, produtos, clientes, pesquisas, trabalhos, entre outras. -, pois a sua análise e processamento é essencial para futuras decisões. São por isso necessárias eficazes técnicas de *data mining*[1] e visualização.

O *data mining,* sendo uma área em constante progressão*,* tem provado ser um processo essencial para as aplicações espaço-temporais, desenvolvendo novas perspectivas para analisar grandes quantidades de informação. Contudo, este não é processo fácil pois tem sido difícil definir técnicas capazes de lidar com os dados para todas as suas interpretações e representações. Na maioria destes sistemas são necessários especialistas de *data mining* para aplicar estas técnicas, e quando surgem resultados são necessários especialistas na área em estudo para tratar esses dados de forma a serem mais facilmente compreensíveis.

Neste contexto, a visualização oferece poderosos meios para descobrir padrões e tendências em dados desconhecidos. A ideia base da exploração de dados visuais é de apresentar os dados de uma forma visualmente agradável, de forma ao utilizador ter uma perspectiva compreensível dos dados, interagir com eles e tirar conclusões. Contudo, uma

[1] Processo projectado para explorar grandes quantidades de dados em busca de padrões

eficaz visualização nem sempre é fácil de atingir, o que já trouxe vários problemas ao longo da história.

A 2 de Janeiro de 1986, o vaivém espacial *Challenger*, explodia 73 segundos após o seu levantamento. A baixa temperatura ambiente, na hora do lançamento, fez com que os anéis de borracha de protecção dos foguetes não funcionassem e começassem a largar gases quentes de encontro ao tanque de combustível e do hardware do foguete, culminando na desintegração do vaivém e consequentemente provocado a morte a toda a tripulação.

Acredita-se que tal desastre podia ter sido evitado, pois houvera várias tentativas por parte de um grupo de engenheiros e gestores para atrasar o lançamento do vaivém, indicando problemas dos anéis de protecção dos propulsores dos foguetes com a temperatura.

A figura 2.3 apresenta um dos gráficos de análise dos danos nos anéis de protecção do propulsor, em lançamentos anteriores, pela qual a decisão do lançamento foi tomada. Neste gráfico, os foguetes estão representados por ordem cronológica de lançamentos. Esta escolha de apresentação ofusca as variáveis de interesse mais importantes. A temperatura é apresentada textualmente em vez de graficamente. O grau de danificação não é apresentado numa escala natural gráfica e nem sequer possui uma legenda. Os diagramas dos foguetes tornam o gráfico desorganizado. De facto, estes erros de representação gráfica inviabilizaram a descoberta de um padrão de problemas. O que se constatou foi que os danos nos propulsores eram relativamente poucos.

Porém, os gráficos das figuras 2.4 e 2.5 mostram os mesmos dados, mas uma história diferente. Através de uma representação mais simples, revelando apenas as duas variáveis de maior interesse, é possível notar-se um claro padrão de danos para lançamentos com temperatura inferior a 65ºF (≈18.3ºC). Em 24 voos antes do *Challenger*, onde a temperatura foi superior a 65ºF, em apenas 4 deles houve problemas com os anéis. Já para temperaturas inferiores a 65ºF houve 4 voos e em todos houve problemas. Acredita-se que tendo tido estes gráficos em vez do apresentado na Figura 2.3, a NASA estaria informada de que era muito arriscado lançar o vaivém naquelas condições, evitando assim o desastre. [18], [19].

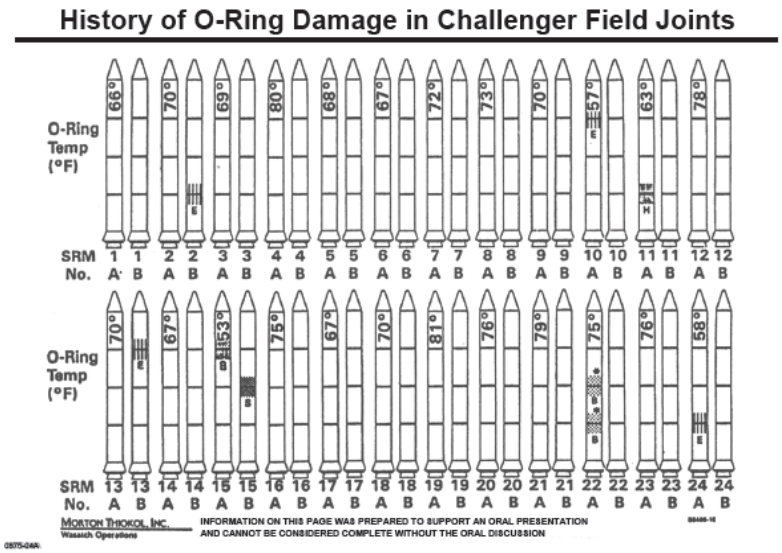


Figura 2.3: Um dos diagramas usados para tomar a decisão de lançar o *Challenger* [18]

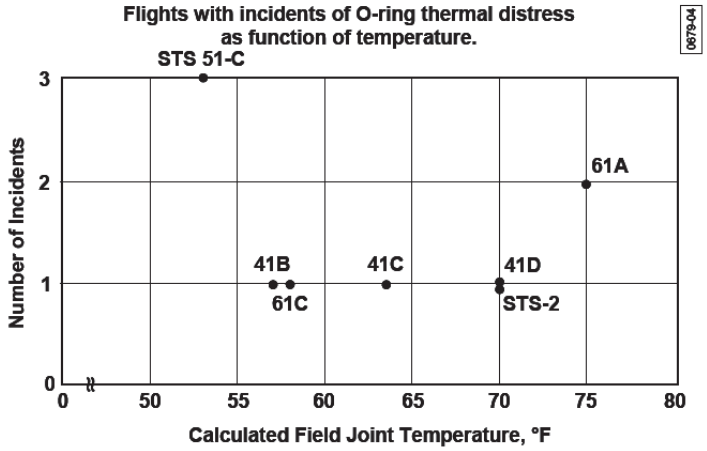


Figura 2.4: Gráfico Incompleto do número de acidentes em função da temperatura [18]

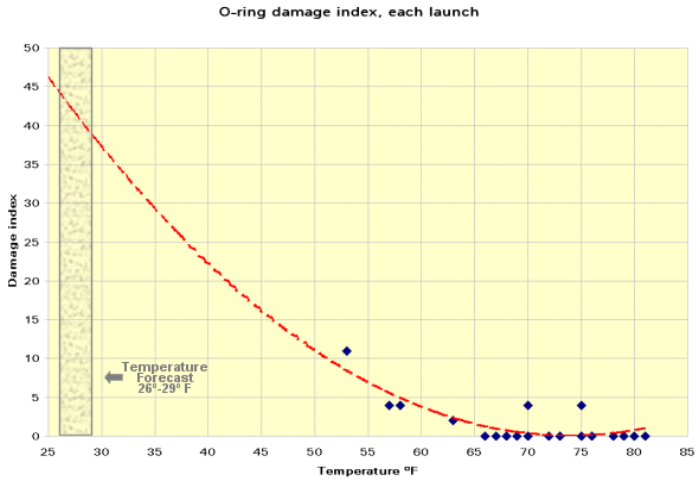


Figura 2.5: Gráfico do número de acidentes por temperatura [19]

Este desastre é um exemplo claro, de que nem sempre é efectuada uma visualização eficiente dos dados. Existe uma grande dificuldade em visualizar grandes quantidades de dados, de aferir o melhor método de representação de forma a contribuir para um conhecimento, principalmente, quando a vida humana está em jogo, como era o caso. Para se devidamente perceber este problema, é fundamental compreender primeiro, a convergência deste, porque surge e de onde surge. Desta forma, é necessário conhecer os diferentes tipos de conteúdo presentes e como se relacionam entre si, nomeadamente a hierarquia Dados – Informação – Conhecimento - Sabedoria (Data – Information – Knowledge - Wisdom). Esta hierarquia é um modelo de classificação da compreensão do ser humano no espaço perceptivo e cognitivo [16].

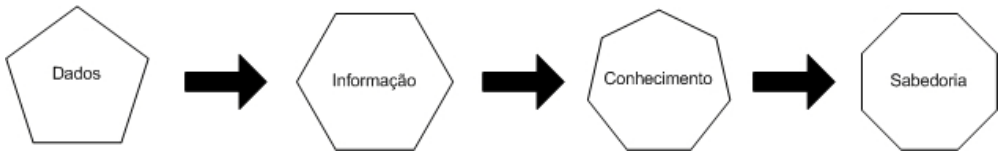


Figura 2.6: Hierarquia Dados – Informação – Conhecimento - Sabedoria

Os Dados podem ser entendidos como uma sequência de símbolos quantificados e qualificáveis. Elementos não interpretados, observações, factos ou características. Não têm significado por si próprios. A palavra Informação possui diferentes significados em diferentes contextos. Contudo, é um termo geral para um tipo fundamental de substância, que é armazenada, processada, transmitida, podendo ter diferentes graus de estruturação. Os Dados após serem processados possuem um significado num determinado contexto. A Informação é o significado que um conjunto de Dados tem para alguém. Este conjunto passa a ser Informação para uma pessoa, quando esta consegue perceber as relações entre os elementos do conjunto que lhe definem um contexto. Por sua vez o Conhecimento é o modelo da realidade que o próprio ser humano dispõe, dentro ou fora da mente. O Conhecimento é uma capacidade, pois é dinâmico. Quem conhece pode estabelecer novas relações, tirar novas conclusões, fazer novas inferências, agregar novas informações, reformular significados. É a Interpretação formal das relações entre Dados e Informação. Factores como a experiência, intuição, raciocínio lógico, podem contribuir para o estabelecimento das relações. Já a Sabedoria é Conhecimento sobre o próprio Conhecimento. O ser humano usa o Conhecimento, avaliando o seu entendimento da realidade, tornando-se capaz de reconhecer se possui Conhecimento completo sobre um assunto, se o mesmo é certo ou errado, bom ou mau, se pode ser usado ou não e em que situações; permite prever tendências e o desenvolvimento de novas teorias. No gráfico da figura 2.7 é possível observar que a complexidade vai aumentando com as interacções e inter-relações necessárias.

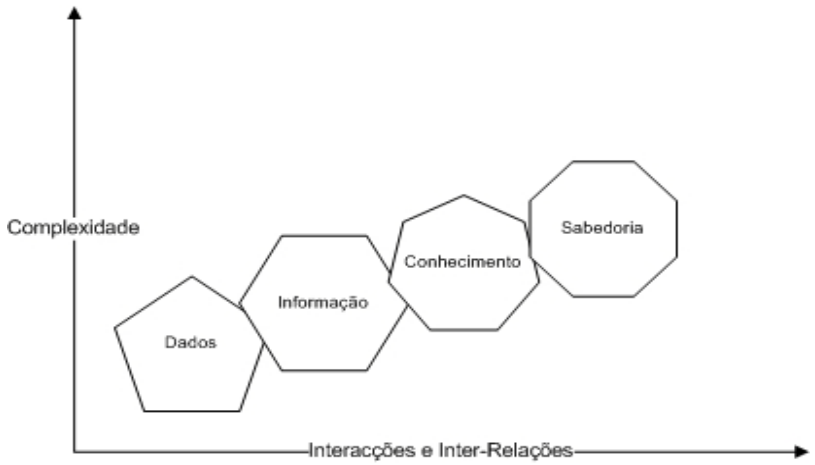


Figura 2.7: Dos Dados à Sabedoria

Em [16], os autores dizem que os Dados são algo menos que a Informação e que a Informação é algo menos que o Conhecimento. Para obter a Informação, é preciso obter os Dados e só depois de se ter Informação é que o Conhecimento pode surgir. Esta Informação necessita de ser comunicada a outros utilizadores na tentativa que estes ao conseguir analisá-la, sejam capazes obter Conhecimento.

2.5.1 - Comunicação da Informação

A comunicação de dados é algo presente em muitos aspectos da vida diária do ser humano. Se o detentor da informação for capaz de explicar a maioria das relações que estabeleceu na aquisição da informação, é eventualmente possível comunica-las a outros, permitindo que estes adquiram informação semelhante. A informação recebida não será exactamente igual à original, pois não será possível garantir que todos os indivíduos retirem o mesmo significado.

Na análise de grandes quantidades de dados, há uma procura por estruturas, padrões, anomalias ou relações entre eles. Esta procura pode ser suportada pela visualização, apresentando os dados em várias formas com variadas interacções. Esta oferece uma visão geral, de uma grande quantidade de dados complexos e assistência para encontrar áreas de interesse e parâmetros para uma análise mais adequada. Várias pesquisas sobre o uso de técnicas de visualização de informação têm sido feitas, para apresentar os dados de um modo gráfico e interactivo de forma a facilitar a sua compreensão. De um modo geral, é possível observar que inúmeras pesquisas procuram representar graficamente o resultado de consultas predefinidas, as quais fornecem resultados que são adequados a situações específicas da análise, o que não permite ao comum utilizador informar que conjuntos de dados são necessários à sua análise. Permitir algo do género, seria extremamente complicado devido à enorme quantidade dos dados e da sua natureza. Por outro lado, aumentaria o trabalho por parte do utilizador para escolher os dados a observar.

Ao longo das décadas, de forma a auxiliar a compreensão, têm sido desenvolvidas formas de visualização, como diagramas e mapas, registando e mostrando dados para serem posteriormente analisados pelas capacidades sensoriais, tendo sido dado maior ênfase à visão. Com esta, o ser humano pode reparar em detalhes e detectar padrões nos dados representados por certas formas de visualização que de outra maneira seriam extremamente difíceis de analisar. O uso da visão para analisar dados permite que o utilizador use a sua percepção visual para os compreender.

O uso do computador facilita, em muito, este tipo de trabalho, principalmente devido ao elevado nível de interactividade que oferece. Sem este, as pessoas seriam forçadas a criar as suas representações, por exemplo, em papel, de acordo com o que elas entendem da tarefa de análise dos observadores. Estas representações seriam estáticas. Os observadores não teriam as opções de interactividade que os computadores permitem, como controlar os dados exibidos num gráfico. As interacções possíveis são definidas pelos projectistas, cabendo a estes ter em mente os possíveis interesses dos potenciais utilizadores de modo a definir quais as interacções que devem ser possibilitadas.

O estudo na área de Visualização de Informação procura optimizar o uso das habilidades visuais do ser humano, facilitando a compreensão dos dados representados visualmente. Procura também revelar visões esclarecedoras sobre fenómenos para os quais não existe uma representação visual espacial inerente.

Uma técnica de visualização é baseada numa representação visual e em mecanismos de interacção que possibilitam ao utilizador manipular essa representação de modo a compreender um conjunto de dados. O nível de abstracção da visualização é geralmente mais elevado, porque o utilizador está mais interessado em observar características ou padrões do que a observar os dados em estado bruto. Uma das principais considerações a ser feita neste processo, é a determinação da técnica que deve ser empregue numa determinada aplicação. Esta escolha depende do tipo de informação que está a ser tratada e das tarefas que precisam

ser realizadas. Em [17], *Shneiderman*, classifica as técnicas de visualização por dados e por tarefas, podendo estas ser unidimensionais, temporais, bidimensionais, tridimensionais, multidimensionais, dirigidas à visualização de hierarquias e de relacionamentos (grafos), e podendo suportar tarefas como *zooming*, filtragem, identificação de relacionamentos ou extracção de várias informações. Em [20] o autor classifica as diferentes técnicas baseando-se em três critérios: dados a serem visualizados, técnica de visualização e a técnica de interacção e distorção usadas. A figura 2.8 mostra essa classificação.

Resumindo, as técnicas de visualização de informação, procuram representar graficamente os dados, de modo a que a representação visual gerada explore a capacidade de percepção do homem e este, a partir das relações espaciais exibidas, as interprete e compreenda para que consiga deduzir novos conhecimentos.

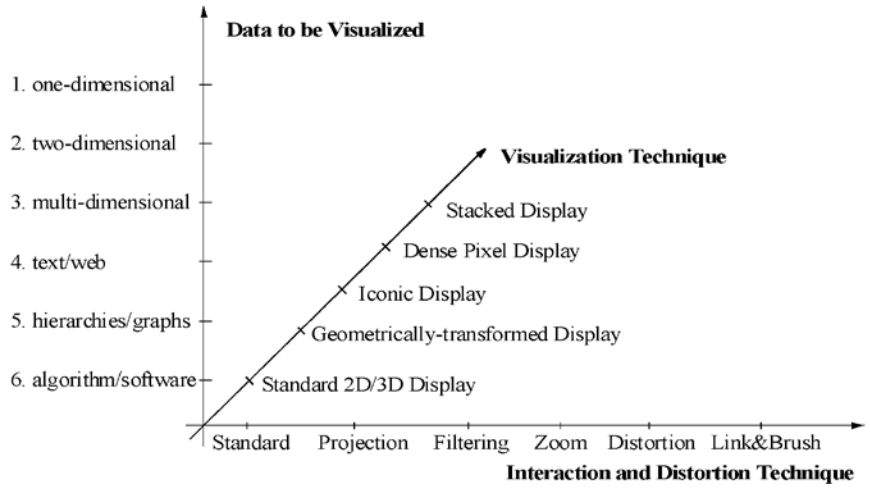


Figura 2.8: Classificação das técnicas de visualização [20]

2.5.2 - Representações Visuais

O avanço das tecnologias tem alterado as formas de visualização dos dados. Por exemplo, nos dados espaço-temporais, os mapas, sendo os tradicionais modos de apresentação, têm vindo a ser complementados por fotos, animações, vídeos, entre outros. As soluções de visualização não devem passar por apenas incluir gráficos estáticos com os resultados produzidos ao longo do processo, mas dar a oportunidade, se possível, de interagir com estes, como por exemplo mudar as posições ou possibilitar *zoom*. Em [10], é dito que nas existentes aplicações geográficas os aspectos de visualização espacial não são adequados para a tomada de decisão quando usados sozinhos. Como tal, desenvolveram um sistema capaz de fazer com que as ferramentas de *data mining* providenciassem alguma forma de localização nos dados que estão a ser analisados e que fosse possível visualizar em 3D, de um modo interactivo, os resultados provenientes desse *data mining*. De referir, que nas ferramentas de visualização usadas, uma explora o *Google Earth* [44] e a outra o *Java3D* [45].

A escolha do tipo de multimédia a usar para esta visualização tem que ter em conta dois aspectos fundamentais: o contexto em que se insere e o que se pretende mostrar ao usa-la. O contexto da aplicação multimédia estará mais orientado para a exploração de dados por especialistas no assunto em questão, de forma a encontrarem respostas a perguntas que

formulem, ou então orientado mais para a demonstração desses dados para pessoas sem grande conhecimento do assunto.

Nesta tese, os dados deverão ser orientados para ambos os casos, ou seja, para que o utilizador comum possa, ao olhar para os dados, entender o que está a visualizar, mas também para que utilizadores instruídos nos assuntos sócio económicos consigam retirar as devidas conclusões.

A visualização de dados espaço-temporais foi e continua a ser, assunto de grande debate, por isso, é possível encontrar um leque variado de material, com opções e critérios para formas de visualização destes dados. Diferentes soluções têm vindo a ser propostas pela ciência *"Visual Analytics"* [8], que pode ser definida do seguinte modo:

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. [Thomas and Cook, 2005] ".

Peter Gatalsky, Natalia Andrienko e *Gennady Andrienko* em [3], fazem uma revisão de algumas ferramentas utilizadas na visualização de dados espaço-temporais, levando em consideração duas perspectivas: a que tipo de dados são aplicáveis e que tarefas de exploração podem suportar.

Duas técnicas de exploração essenciais nas ferramentas de visualização, baseadas em ambiente informático, são a possibilidade de *querying* e a animação. O *querying* permite que o programa responda às perguntas levantadas pelos utilizadores sobre os dados espaço-temporais. Por outro lado, a animação é uma grande ajuda na visualização de certos dados para, por exemplo, encontrar padrões que, sem animação, se tornam muito complicados de serem verificados mesmo pelos especialistas. Podem também servir para tornar a visualização mais agradável para o utilizador comum. As alterações nos dados são representadas por alterações no ecrã. É possível encontrar animações em grande parte dos programas que lidam com dados espaço-temporais.

É comum ser possível aceder aos dados de objectos espaciais ou localizações através de um mapa ou outro tipo de gráfico. O utilizador apenas precisa de orientar o cursor sobre esse objecto ou localização. O posicionamento deste indica os resultados das *queries* feitas por essa posição. Há que estudar as variáveis em questão, as projecções para visualização e também ter em conta o tamanho destas, *zoom*, entre outros aspectos.

Em análise gráfica, é costume ter-se *queries* gráficas, por exemplo, colorir determinada localização ou objecto, de forma a salientar determinado acontecimento. De forma a incrementar a percepção, pode também ser disponibilizada outra perspectiva dessa visualização.

A natureza diversificada das bases de dados espaço-temporais sejam científicas, sociais, estatísticas, ou de outro tipo, faz com que estas necessitem de um elevado número de ferramentas criativas, com um considerável nível de desempenho e interactividade.

Relativamente às técnicas de visualização, existem inúmeras opções de representação de dados. Algumas das técnicas mais simples e comuns são os gráficos X-Y, X-Y-Z, gráficos de colunas, gráficos de linhas, gráficos de áreas, gráficos circulares e mapas simples. De um modo geral, estas técnicas apenas são capazes de representar no máximo duas ou três dimensões, dependendo dos casos.

Dependendo da aplicação, existe um elevado número de técnicas que não sendo tão simples podem ser muito mais eficazes. Neste contexto, destacam-se as Representações a 3D, *Treemaps, Cone Trees*, Coordenadas Paralelas, Diagramas de *Venn*, *Hyperbolic Trees, Cluster Diagrams*, Mapas complexos, *Iconic Displays*, entre outras. Com algumas destas técnicas já se

torna possível representar três ou mais dimensões. De forma a colmatar as necessidades dos utilizadores, pode-se ainda, aplicar várias técnicas na mesma visualização.

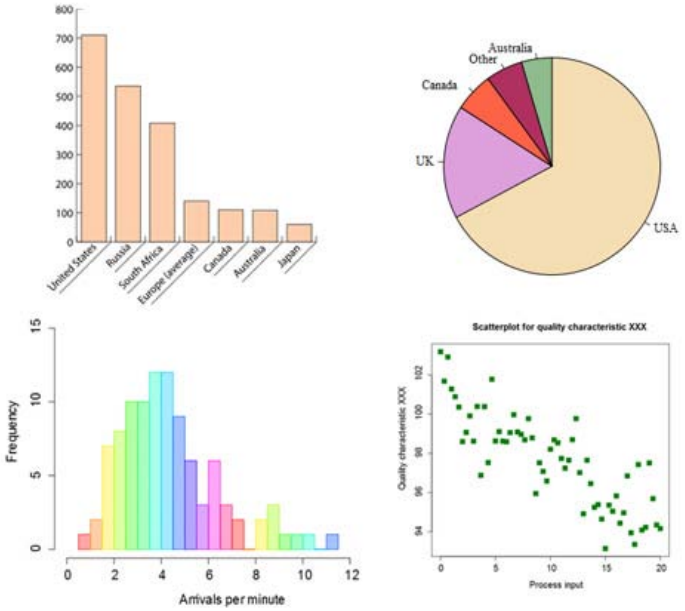


Figura 2.9: Alguns exemplos de visualização. Da esquerda para a direita: Em cima, gráfico de colunas e gráfico circular. Em baixo, gráfico de dispersão e histograma, [54]

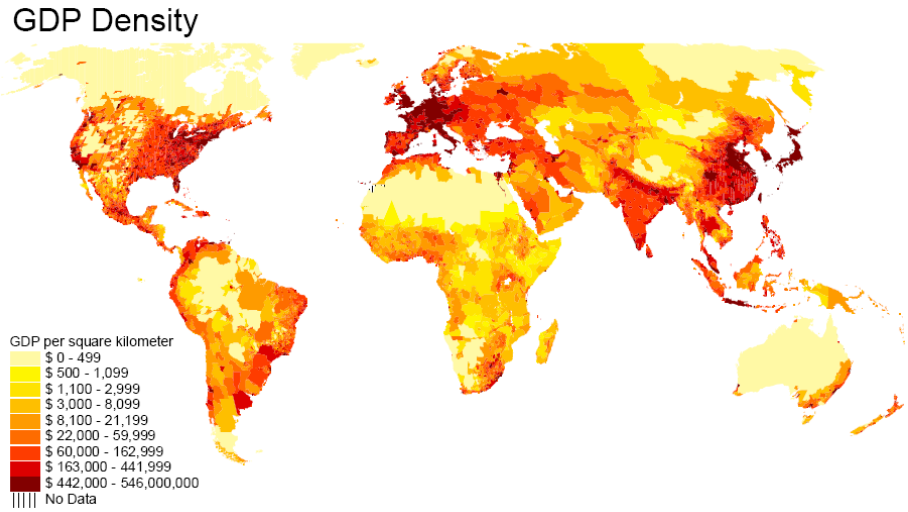


Figura 2.10: Representação em mapa. As diferentes cores representam diferentes atributos nos espaços, [47]

Figura 2.11: *Treemap* [54]

Para a representação dos dados, existem inúmeras vantagens em utilizar-se tipos de visualização mais simples, já existentes. Para além de já terem dado provas de funcionarem para específicos conjuntos de dados, o grau de familiarização dos utilizadores é maior.

Contudo, a escolha da técnica a usar pode ser árdua, devido ao elevado leque de opções. É necessário ter em conta, que não existem técnicas, que sejam aplicáveis ou que funcionem eficientemente para todos os tipos de dados.

Em [14] é proposta uma forma de abordagem aos diferentes métodos de visualização de dados temporais. Esta abordagem é baseada em três critérios principais: tempo, dados e representação. Os autores dividem a representação em dois subcritérios: dependência do tempo, isto é, estático ou dinâmico, e em dimensionalidade, 2D ou 3D. A maioria dos fenómenos no mundo real são dinâmicos, como tal, o termo estático só pode ser usado para objectos que não sejam alterados num período curto de tempo como mapas cartográficos, estradas, edifícios, entre outros. Por outro lado, a informação dinâmica refere-se à informação sobre os objectos geoespaciais que variam num curto período de tempo, cujo comprimento será definido de acordo com os parâmetros em uso. O facto de se optar pela visualização 2D ou 3D recai sobretudo sobre a sua usabilidade e o grau de complexidade que se pretende. Por um lado, o 2D torna as visualizações mais acessíveis, por outro com o 3D permite englobar uma maior quantidade de informação. Por exemplo, na visualização de dados volumétricos, será mais adequado recorrer-se ao 3D.

Infelizmente, embora existam diversos tipos de visualização de dados, nem sempre os requisitos são satisfeitos. Muitos conjuntos de dados têm características únicas que forçam os utilizadores a surgirem com novas formas de representação.

Na literatura de referência, existe um elevado número de novas ideias de visualização de dados. A representação de dados que variam no tempo e no espaço é um assunto de grande debate, o que faz com que muitos autores falem sobre novas formas de visualização. Em [9] os autores, escrevem sobre a implementação de um cubo espaço-temporal para a representação de dados. De acordo com os autores, este cubo surge por sentirem a falta de uma ferramenta capaz de representar os dados espaço-temporais de uma forma continua. Este é usado para representar dados sobre os movimentos espaciais e as interacções espaciais, isto é, alterações nas propriedades espaciais nos objectos discretos. Estes são

representados como círculos colocados verticalmente de acordo com o tempo de ocorrência. Na figura 2.12, é possível ver um exemplo deste cubo.

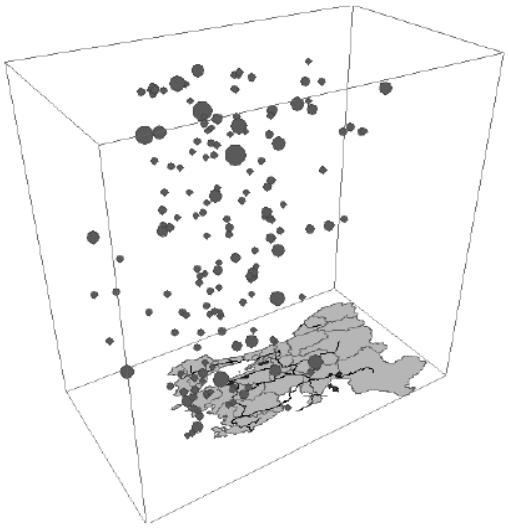


Figura 2.12: "*Space-Time Cube*", [9]

Os eventos iniciais estão situados na base do cubo e os mais recentes no topo. Para representar factores associados aos eventos são usadas características gráficas como áreas, cores, entre outras. Os autores tornaram esta representação bastante dinâmica, podendo o utilizador alterar o campo de visualização, modificando o período de tempo a observar e a perspectiva de visão, efectuando *zoom*, entre outras possibilidades. O cubo foi usado neste artigo, para analisar os sismos em *Marmara*, *Turquia*.

Em [21], os autores descrevem uma forma de visualização de dados espaço-temporais em mapas. O conceito passa por introduzir ícones 3D num mapa para a representação dos dados espaço-temporais, e na introdução de métodos baseados em eventos para reduzir a quantidade de informação a ser apresentada, figura 2.13.

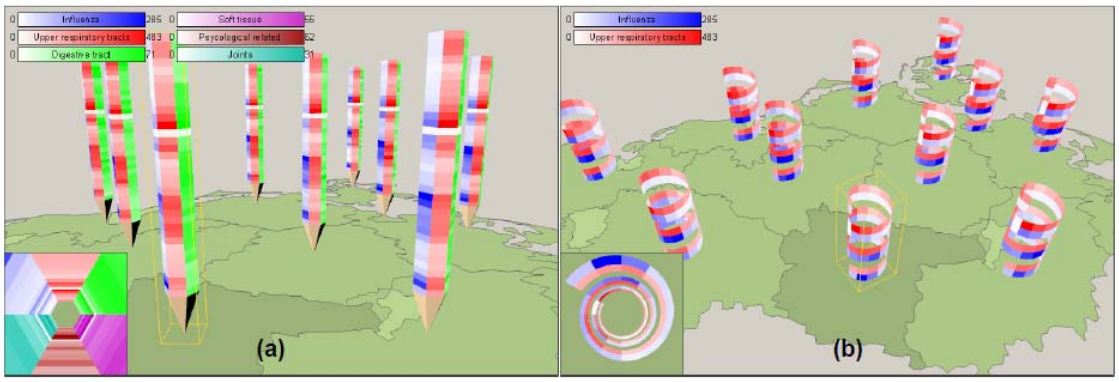


Figura 2.13: Representação da saúde humana com Ícones em mapas [21]

As duas formas de visualização supra descritas, podiam ser usadas para representar vários casos do Observatório. Por exemplo, poderiam ser aplicadas na visualização da variação do desemprego, ao longo do tempo, nas diferentes freguesias do município de Santa Maria da

Feira. Porém, a sua implementação no Observatório não seria exequível, pois iria ser extremamente complicado para os utilizadores compreenderem essas formas de visualização, para as quais não possuem formação (formal e/ou informal). Para além disso, estas estratégias de visualização teriam que ser suportadas pelos programas de navegação, o que limita a sua concepção.

Um autor que também estuda casos de visualização complexa é *Manuel Lima* [39]. Na sua página pessoal é possível encontrar vários destes exemplos. Visto que a informação pública em Portugal é livre mas geralmente apresentada em formas que não permitem a compreensão de padrões relevantes, o autor através de visualizações extremamente ricas, tenta realçar factos em vários conjuntos de dados. A figura 2.14 representa uma dessas visualizações. Com um gráfico muito interactivo, o autor ilustra a data de criação de 308 municípios portugueses ao longo de um milénio. É possível interagir com as colunas que representam as décadas e os com os círculos que representam os séculos. É também possível interagir com o mapa para se ter uma visão sobre o ano de criação de cada município. A visualização possui também um botão de *"Play"*, que inicia uma animação para se perceber melhor como Portugal foi sendo formado.

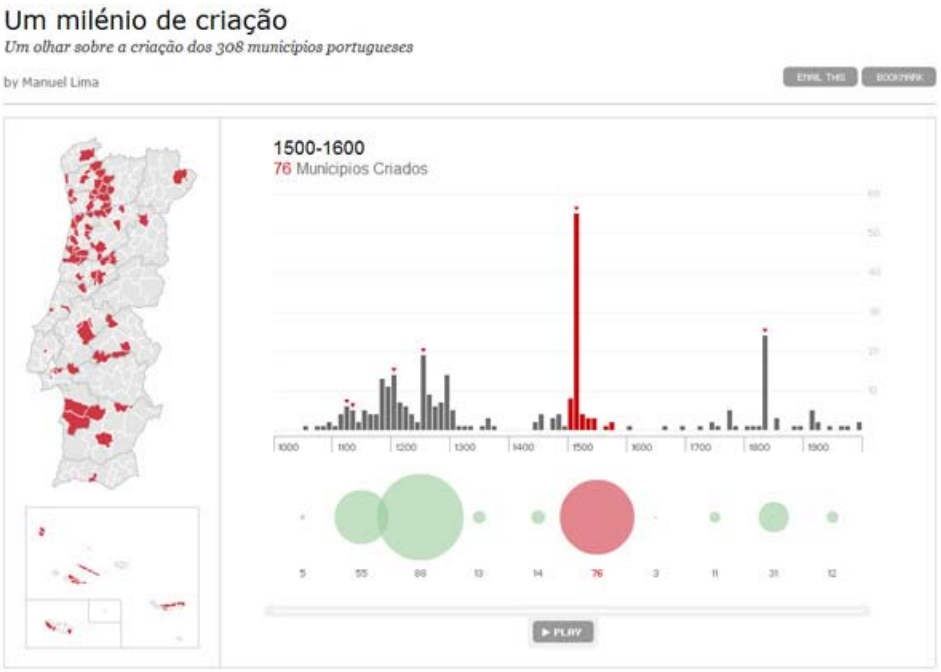


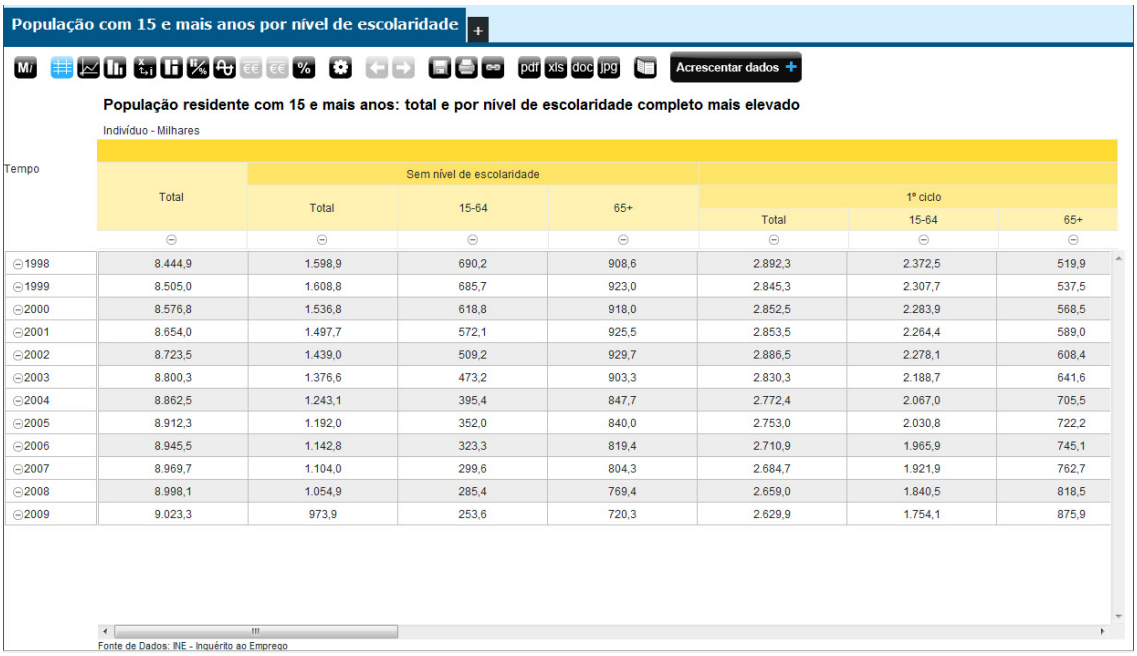
Figura 2.14: Um milénio de criação [39]

De facto, existem bastantes soluções espalhadas pela internet sobre a visualização de dados espaço-temporais ou dados multidimensionais. Contudo, nem sempre são boas soluções ou aplicáveis no âmbito deste trabalho. Um dos exemplos é *DCcrimeViz* [40], que recorre ao uso da API do G*oogle Maps* [52] para a visualização dos crimes ocorridos, figura 2.15. O utilizador pode navegar dinamicamente pelos diferentes intervalos de tempo, com as colunas situadas no fundo da imagem, sendo apresentados no mapa os diferentes locais onde os crimes ocorreram. A sua visualização é extremamente dinâmica, pois possui várias funcionalidades para o utilizador controlar a sua visualização, contudo torna-se pouco eficaz na análise de padrões ou na formulação de conclusões.

Figura 2.15: Exemplo de visualização no *DCcrimeViz*, [40]

Analisando as representações visuais de outros observatórios, é possível constatar que muitas não são eficazes, como será o caso do Pordata [48]. O Pordata é uma base de dados que contem vários dados sobre Portugal. É constituído por um vasto leque de temas e, tal como o Observatório Social de Santa Maria da Feira, apresenta os dados em tabelas e gráficos. Possui uma plataforma própria, com um vasto leque de opções, permitindo ao utilizador filtrar os dados que pretende observar, o que a torna bastante dinâmica e interactiva. Na figura 2.16 temos um exemplo de apresentação de dados em tabela no Pordata. Estes dados são referentes à "População com 15 e mais anos por nível de escolaridade". O Pordata faz uso da *ScrollBar* para permitir ler a tabela de uma forma contínua.

Tal como o Observatório Social desta dissertação, o Pordata, ainda padece de alguns problemas de apresentação gráfica de dados, apresentando duas estratégias diferentes. A primeira apresentação é feita através de um gráfico de linhas, tornando-se muito confusa por possuir várias variáveis, figura 2.17. Não é possível olhando para o gráfico perceber o que cada linha representa, apesar de possuir legenda. As cores também são muito parecidas, o que torna tudo mais confuso. É demasiada informação num só gráfico para se conseguir a sua compreensão. Fazendo uso da API da *Google,* a segunda apresentação revela-se mais dinâmica. Dentro dessa apresentação é possível observar os dados de três maneiras diferentes: gráficos de colunas, figura 2.18, de linhas, figura 2.19, ou de círculos, figura 2.20. Possui também um botão *"Play"* para apresentar a sua evolução ao longo do tempo. No entanto, esta apresentação revela-se pobre, pois como é possível observar pelas figuras, apenas revela uma pequena parte dos dados e só se consegue efectuar uma análise precisa se se comparar os dados da tabela com os do gráfico.



População com 15 e mais anos por nível de escolaridade

Acrescentar dados

População residente com 15 e mais anos: total e por nível de escolaridade completo mais elevado

Indivíduo - Milhares

Tempo	Total	Sem nível de escolaridade			1º ciclo		
		Total	15-64	65+	Total	15-64	65+
1998	8.444,9	1.598,9	690,2	908,6	2.892,3	2.372,5	519,9
1999	8.505,0	1.608,8	685,7	923,0	2.845,3	2.307,7	537,5
2000	8.576,8	1.536,8	618,8	918,0	2.852,5	2.283,9	568,5
2001	8.654,0	1.497,7	572,1	925,5	2.853,5	2.264,4	589,0
2002	8.723,5	1.439,0	509,2	929,7	2.886,5	2.278,1	608,4
2003	8.800,3	1.376,6	473,2	903,3	2.830,3	2.188,7	641,6
2004	8.862,5	1.243,1	395,4	847,7	2.772,4	2.067,0	705,5
2005	8.912,3	1.192,0	352,0	840,0	2.753,0	2.030,8	722,2
2006	8.945,5	1.142,8	323,3	819,4	2.710,9	1.965,9	745,1
2007	8.969,7	1.104,0	299,6	804,3	2.684,7	1.921,9	762,7
2008	8.998,1	1.054,9	285,4	769,4	2.659,0	1.840,5	818,5
2009	9.023,3	973,9	253,6	720,3	2.629,9	1.754,1	875,9

Fonte de Dados: INE - Inquérito ao Emprego

Figura 2.16: Exemplo da plataforma de visualização de dados do Pordata [48]

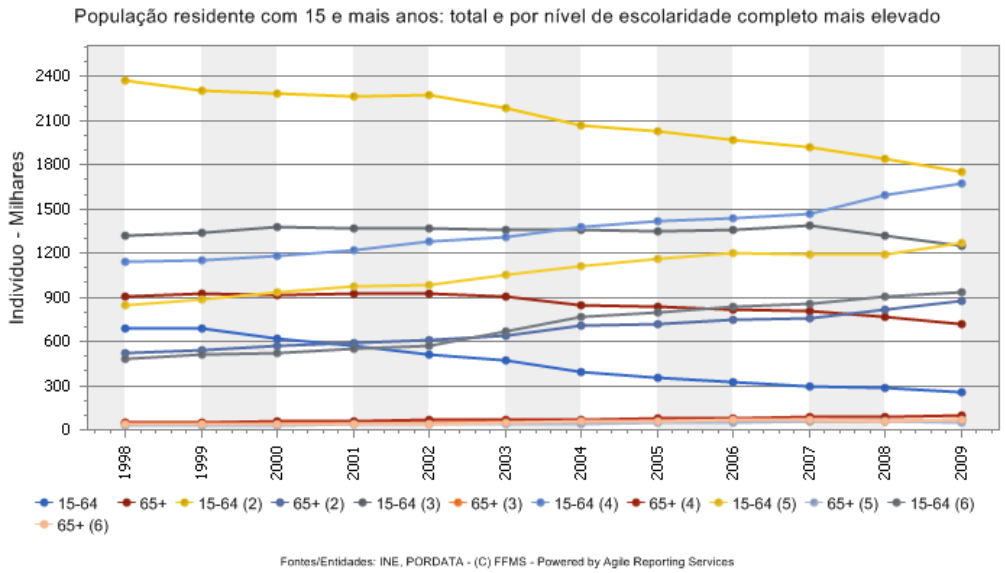


Figura 2.17: Primeira forma de visualização gráfica da tabela da figura 2.16 [48]

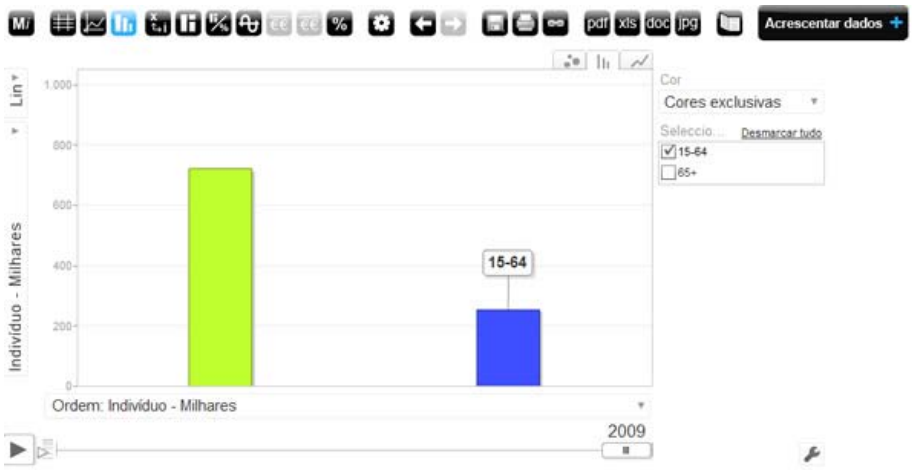


Figura 2.18: Segunda forma de visualização gráfica da tabela da figura 2.16: gráfico de colunas [48]

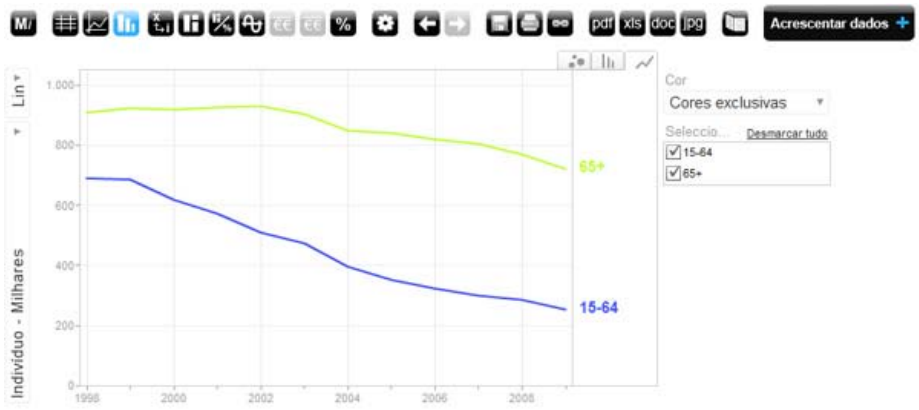


Figura 2.19: Segunda forma de visualização gráfica da tabela da figura 2.16: gráfico de linhas [48]

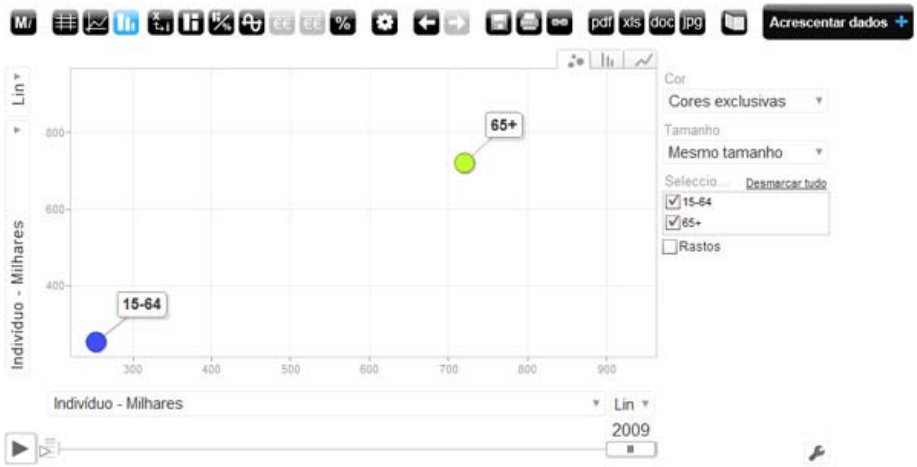


Figura 2.20: Segunda forma de visualização gráfica da tabela da figura 216: gráfico de bolas [48]

Como é possível constatar, quando existem mais do que duas dimensões em acção, é extremamente difícil apresentar grandes quantidades de dados. Efectivamente, diversos portais de estatística preferem apenas mostrar os dados em tabelas ou apenas casos simples com uma ou duas dimensões.

Este é um dos grandes desafios a que se tentou responder.

2.6 - Padrões de Utilização

De modo a melhorar as estruturas e os conteúdos dos *Websites*, existe actualmente uma crescente necessidade de analisar como os utilizadores interagem com estes. De um modo geral, o conteúdo e a estrutura de um *Website* estão destinados a um determinado propósito. Contudo, o comportamento dos utilizadores pode revelar relações ou estruturas que não foram pensadas ou projectadas anteriormente. O *Web Mining* é a aplicação de técnicas de *Data Mining* em páginas *Web,* de forma a extrair padrões de utilização.

Em muitos casos, as estruturas de um *Website* são extremamente complicadas causando muitas vezes com que os utilizadores falhem os objectivos que pretendem atingir quando o visitam. O *Web Mining* permite analisar os utilizadores de forma a melhorar os *Websites* e evitar este tipo de acontecimentos. Por outro lado, o *Web Mining* pode também ajudar a antecipar as necessidades dos utilizadores, o que em termos de negócio e comércio pode ser fundamental.

Para a descoberta de conhecimento numa base de dados, é indispensável ter a noção do conjunto de dados a que se pretende aplicar as técnicas de *data mining*. Os dados podem ser recolhidos ao nível do servidor, do cliente ou em servidores *proxy*.

A recolha ao nível do servidor é muito importante, pois regista a informação sobre a navegação dos vários utilizadores, embora estes registos possam não ser totalmente fiáveis devido à presença de vários níveis de *caching* na *Web;* páginas em *cache* não são guardadas num registo de servidor. Um servidor *Web* pode também guardar outros tipos de informação de utilização, como por exemplo os denominados *cookies*. Numa simples definição, os *cookies* são dados gerados pelo servidor sendo gravados localmente pelos programas de navegação dos utilizadores; representam parâmetros que permitem identificar um utilizador durante as suas navegações pelo *Website*. Os conteúdos de um *cookie* variam de acordo com os *Websites*. Geralmente é gravada a identificação e a palavra-chave dos utilizadores, podendo também ser gravadas informações adicionais como actividades no *Website*, compras, cartões de crédito etc. Contudo os *cookies* estão dependentes da aceitação dos utilizadores.

A recolha ao nível do cliente pode ser implementada usando um agente remoto como JavaScript, ou modificando o código-fonte de um *software* de navegação como o *Mozilla*, de forma a aumentar as suas capacidades de recolha de dados. Este nível de recolha apenas funciona com a cooperação dos utilizadores, activando o JavaScript ou usando os programas de navegação modificados.

Estas recolhas estão sujeitas ao tipo de dados disponíveis, ao segmento de população onde são recolhidos e ao seu método de implementação. Em [25], os autores classificam os dados que podem ser usados em *Web Mining* em quatro tipos: *Content, Structure, Usage* e *User Profile*. Os dados *Content* são aqueles cuja página *Web* foi concebida para apresentar aos utilizadores; os *Structure* são dados que descrevem a organização do conteúdo, como por exemplo entidades de dados como HTML ou *tags* XML; os *Usage* são dados que descrevem os padrões de uso das páginas *Web*, como endereços IP, referências das páginas e a data e hora

de acesso, percursos, entre outros; *User Profile* são os dados que fornecem informação sobre os utilizadores, incluindo informação demográfica do utilizador como nome, idade, país, nível de escolaridade, entre outros. Esta informação pode ser obtida através de formulários preenchidos pelo utilizador.

2.6.1 - *Web Usage Mining* e Análise de Registos

O *Web Usage Mining* tem como objectivo revelar conhecimento em registos de utilização de um determinado *Website*. Aplicando técnicas de *data mining* e análises estatísticas, pretende extrair relações entre páginas, padrões de navegação, ou outro tipo de informação de um *Website* e dos seus utilizadores. A figura 2.21 mostra um simples exemplo de como funciona o processo desde a extracção dos Dados até ao Conhecimento.

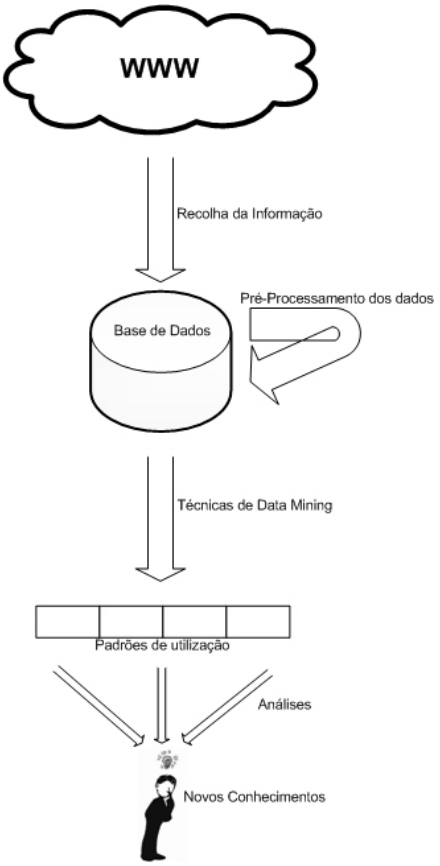


Figura 2.21: *Data Mining* na *Web*

De forma a contribuir para a coerência do *Web Usage Mining*, o *W3C Web Characterization Activity* [51] publicou um esboço de algumas definições de termos *Web* relevantes para a análise da utilização. *Users* (Utilizadores), *Page Views* (Acessos), *Click-Streams* (Sequência de Cliques) e *Server Sessions* (Sessões de Servidor). Um *user* é definido como uma pessoa singular que está a aceder a um servidor Web através de um software de navegação. Uma *page view* é definida como o pedido de consulta de uma página de um *Website*, podendo esta ser formada por qualquer quantidade de ficheiros. Resulta de um clique numa hiperligação. Um *click-stream* é uma sequência de cliques realizados por um utilizador. Já uma *server session* é o *click-stream* para uma visita única de um utilizador num servidor.

De acordo com [26] existem três tarefas principais para se efectuar *Web Usage Mining* ou *Web Usage Analysis*: pré-processamento de utilização, descoberta de padrões e análise de padrões.

O pré-processamento consiste em converter a informação de utilização, conteúdo e estrutura nos termos *Web* definidos pela *W3C Web Characterization Activity*. Tal como foi referido anteriormente, a identificação dos utilizadores é extremamente complicada sobretudo se for feita no lado do servidor. Existe a possibilidade de ocorrem diversos erros se apenas se considerar o endereço IP, o software de navegação e os *click-streams* como relevantes para a sua identificação. Se existir um único endereço IP para vários utilizadores ou então múltiplos endereços IP para apenas um utilizador, a informação é claramente insuficiente. Existem outros métodos para se obter a identificação dos utilizadores. Um deles é o dos já referidos *cookies*, embora também sofram de outros problemas para além de estarem dependentes da aceitação dos utilizadores. Existe a possibilidade do utilizador apagar os *cookies* o que fará com que num novo acesso ao *Website* seja considerado um novo utilizador. Para além disto, a utilização exclusiva de *cookies* não permite a distinção de diferentes utilizadores de um mesmo computador. Outro dos métodos para identificar utilizadores é a utilização de pré-registos online de forma a obter informação relevante como o nome, idade, escolaridade etc. Tal informação é gravada na base de dados e sempre que o utilizador se autentica no sistema, o histórico de navegação fica-lhe automaticamente associado. Contudo, também não é uma solução perfeita, pois os utilizadores podem criar registos falsos, podem criar múltiplos registos, ou vários utilizadores podem-se servir de um único registo.

O pré-processamento de conteúdo, é possível afirmar que este consiste em efectuar classificações ou aglomerações do diferente conteúdo de um *Website*. Pode ser aplicado para filtrar as entradas ou saídas dos algoritmos de descoberta de padrões. Limitar os resultados de padrões àqueles que correspondem a um determinado tema ou determinada área é um exemplo de como este pré-processamento pode ser efectuado.

O pré-processamento da estrutura é feito de modo idêntico ao pré-processamento de conteúdo.

Podem ser feitas várias análises de forma a extrair conhecimento sobre os utilizadores para um variado tipo de aplicações. As mais comuns são as análises estatísticas. Analisando os diferentes registos torna-se possível verificar quais as páginas mais visualizadas, os tempos de visualização por página, os comprimentos de percursos, etc. Esta informação pode ser bastante importante para a melhoria do *Website* ou como suporte a acções sociais ou de mercado. É também costume aplicar regras de associação por forma a relacionar páginas que, por possuírem uma estrutura ou conteúdos idênticos, são geralmente referenciadas em conjunto numa sessão. A aplicação de algoritmos deste tipo pode revelar relações entre utilizadores que visitaram páginas diferentes, o que pode ser muito útil para aplicações de negócios ou marketing. Existem outros tipos de análise que poderão ser feitas, mas tudo depende da finalidade do sistema e dos níveis de conhecimento dos seus responsáveis. Alguns exemplos podem ser consultados em artigos como [27], [28], [29] e [30].

O grande objectivo de qualquer *data mining* está em fornecer aos analistas resultados que sejam interessantes e relevantes para o caso em estudo.

Muitas vezes a estrutura do endereço de um *Website* revela por si só um grafo direccional. O grande problema está na qualificação do conteúdo das páginas. Deve, por isso, ser sempre realizado um pré-processamento do endereço. Para um *Website* estático, a tradução do

endereço é relativamente simples pois cada endereço representa conteúdos diferentes. Já para um *Website* dinâmico a tradução do endereço pode ser extremamente complicada, pois vários endereços podem apontar para o mesmo conteúdo e vários fragmentos de conteúdo podem possuir o mesmo endereço. Um simples exemplo é apresentado na tabela 1.

Tabela 2.1: Exemplos de tradução de endereços

Endereços	1- **www.exemplo.pt/exemplo?id=21&ref=calcado&dest=pagina1** 2- **www.exemplo.pt/exemplo?id=21&ref=calcado&dest=produto&produto=nike23123** 3- **www.exemplo.pt/exemplo?id=33&ref=calcado&dest=pagina1** 4- **www.exemplo.pt/exemplo?id=33&ref=calcado$dest=produto&produto=nike23**
Expressões Regulares	..dest=\(.*)\ ..dest=produto&produto=\(.*)\
Traduções	1- pagina1 2- nike23123 3- pagina1 4- nike23

Apesar de apresentar quatro endereços diferentes, o 1º e o 3º traduzem-se na mesma página. Neste caso são usadas apenas duas expressões regulares[2] mas muitos *Websites* necessitam de elevados números de expressões para traduzir todas as páginas possíveis.

Se um único endereço apontar para diversos conteúdos, os percursos dos utilizadores serão impossíveis de registar pelo lado do servidor. Isto acontece, por exemplo, quando as páginas são formadas por pedidos POST. Um endereço indica como uma página vai ser requisitada a um servidor. O mais comum para efectuar os pedidos é usar-se o GET, deste modo já é possível registar os percursos. A figura 2.22 mostra o exemplo de um simples *Website* de desporto.

Muitas ferramentas de análise de tráfego *Web* produzem periodicamente registos que contêm várias estatísticas sobre os utilizadores e o modo como estes interagem com o servidor. O objectivo será escolher uma que satisfaça as necessidades de análise desta dissertação.

[2] Uma expressão regular descreve um conjunto de cadeias de caracteres, de forma sucinta, sem necessitar de listar todos os elementos do conjunto

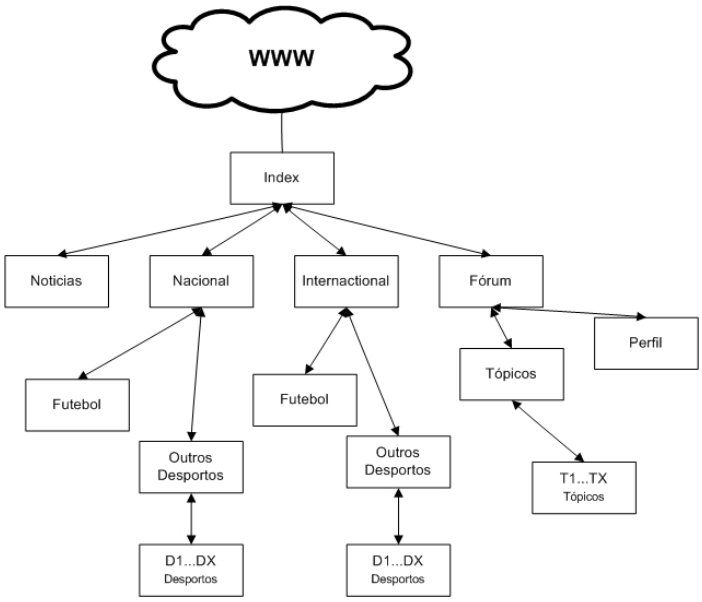


Figura 2.22: Exemplo de um *Website* de Desporto

2.6.2 - Processamento dos dados

Dependendo dos servidores e das aplicações que estes usem para criar os registos de utilização, será sempre necessário efectuar um processamento dos dados de forma a os poder usar. E sendo os dados registados de uma forma automática e pré-definida pelas aplicações, a maior parte das vezes os utilizadores necessitam de expurgar a parte de informação que não interessa e formatá-los de acordo com as necessidades. Isto depende também dos intervenientes e do que estes pretendem fazer com os próprios dados. As ferramentas de análise de tráfego extraem os dados brutos dos registos *Web* e processam-nos de forma a obter informação válida, como o número de visitas a páginas, páginas mais vistas, páginas de entrada e saída, identificação de utilizadores, programas de navegação usados, *cookies*, entre outros. Algumas ferramentas ainda fornecem informações extra, que, entre outras, incluem percursos usados pelos utilizadores e tentativas de entradas não autorizadas.

Capítulo 3

Metodologia

Neste capítulo será descrito o modo de funcionamento do Observatório Social de Santa Maria da Feira. Como é frequente acontecer numa primeira versão de uma aplicação ou software, há problemas que apenas foram detectados após a implementação do sistema com recurso a situações reais. Por isso, antes de passar à concretização dos objectivos do presente trabalho, foi necessário resolver alguns problemas mais críticos, inerentes ao desenvolvimento. Neste capítulo descrever-se-ão esses problemas e os passos efectuados para se chegar à implementação final, nomeadamente opções tomadas e tecnologias escolhidas.

3.1 - Introdução

De modo a iniciar-se o trabalho, foi necessário efectuar um estudo aprofundado de todo o Observatório.

Com uma breve análise ao Observatório e considerando o que foi dito nos capítulos anteriores, é possível salientar quatro importantes componentes em que o trabalho desta tese vai incidir: Armazenamento de Dados, *Query*, Visualização e Análise. A figura 3.1 mostra como estes componentes estão relacionados. As *queries* (*Query*) seleccionam os dados do local de armazenamento de modo a serem visualizados (Visualização). As *queries* serão também o motor para a análise.

Numa primeira fase foi estudada a sua estrutura, incluindo a sua base de dados. O Observatório foi criado com as linguagens PHP, HTML, JavaScript e SQL.

- **PHP** – Linguagem de programação utilizada para gerar conteúdo dinâmico na WWW. Foi criada por *Rasmus Lerdof*, por volta de 1995 com o nome de *Personal Home Page Tools*. Em 1997, *Zeev Suraski* e *Andi Gutmans* desenvolveram o PHP 3, a primeira versão que mais se aproxima do PHP usado. Actualmente a versão usada é o PHP 5. Lançada em 2004, melhorou a orientação a objectos através da introdução de um novo modelo. É ideal para o uso de servidores *Web* pois simplifica a interacção das páginas com as bases de dados.
- **HTML** – Linguagem desenvolvida por *Tim Berners-Lee* em 1990, usada para criar páginas *Web*. Possibilita apresentar informação através da internet. É de simples

construção e aprendizagem, devido à abundante documentação e tutoriais. Possui normalização o que garante suporte dos vários programas de navegação.

- **JavaScript** – Linguagem orientada a objectos muito usada em páginas *Web* no lado do cliente tornando a apresentação e interactividade das páginas *Web* muito superiores. Foi criada pela *Netscape* em 1995. Apesar de semelhante com Java é diferente no seu uso, sendo o JavaScript mais adequado para páginas *Web*. Os programas de navegação são os elementos responsáveis por interpretar e executar os *scripts*, sem ter que apelar aos recursos do servidor, aumentando a rapidez do processo.
- **SQL** – Linguagem que permite gerir os dados de uma base de dados, originalmente baseada em álgebra relacional. Foi desenvolvida na *IBM* por *Donald D. Chamberlin* e *Raymond F. Boyce* no inicio dos anos 70.

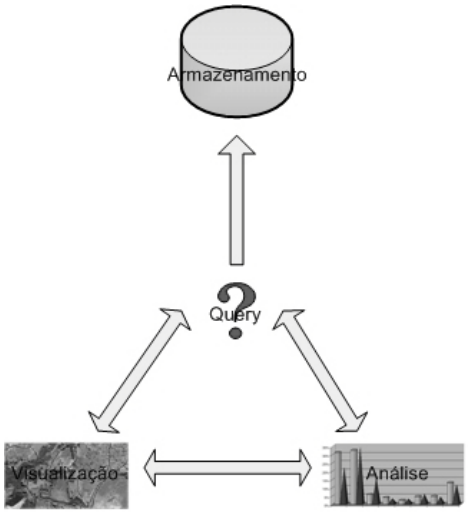


Figura 3.1: Principais componentes no projecto

Estruturar um observatório nem sempre é tarefa fácil. Este necessita de guardar grandes quantidades de informação e gerar as respectivas representações visuais, que podem ser tabelas, gráficos ou outros tipos de visualização mais dinâmicos. Tendo que ser capazes de suportar grandes quantidades de informação referente aos mais variados assuntos, um observatório necessita de estar preparado para fazer as sinapses entre dados. Esta necessidade é extremamente complexa de implementar e nem sempre é possível encontrar a melhor solução. É extremamente importante na sua construção, considerar qual o tipo mais adequado de representação dos dados, assim como o problema mais complexo que vai albergar, isto é, no problema de associação de dados mais complexo que vai ter que representar nas formas de visualização que escolheu.

Na figura 3.2 é possível encontrar um diagrama simplificado da estrutura visível ao utilizador do Observatório Social de Santa Maria da Feira. O utilizador comum possui uma interface alheia à inserção de dados e pode navegar entre as várias Áreas do Observatório Social que, de momento, são cinco - Educação, Emprego, Família, População e Comportamentos Desviantes. Cada uma destas áreas contem temas, e através dos temas o utilizador tem acesso aos indicadores. Os indicadores estão divididos por grupos de acordo com o que representam. Os indicadores são os responsáveis pela visualização dos dados, sendo muito provavelmente a parte mais complexa na construção de um observatório.

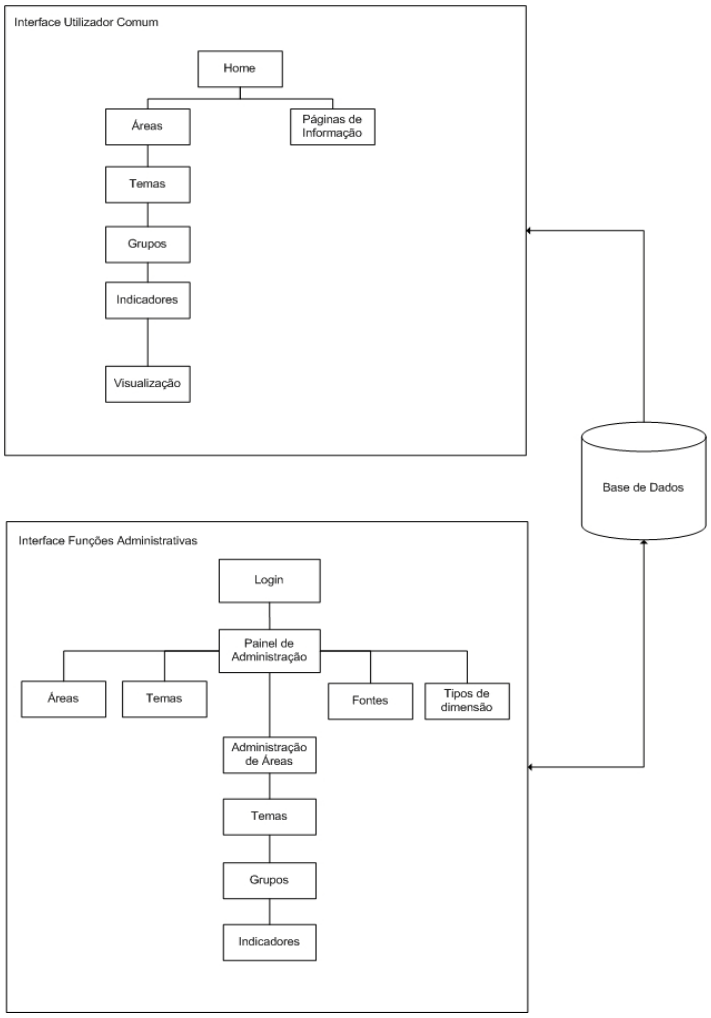


Figura 3.2: Estrutura visível ao utilizador do Observatório Social de Santa Maria da Feira

Toda a informação é registada na base de dados em função da estrutura acima identificada. A base de dados tem que estar de tal forma organizada para que a enorme quantidade de dados seja visualizada sempre no local certo. Grande parte da complexidade prende-se com a estrutura de armazenamento e visualização dos dados nos indicadores. As associações de informação podem estar representadas em duas, três ou até quatro dimensões. Foi necessário pensar o observatório de forma a que estas associações fossem possíveis de representar nas tabelas e nos gráficos correctos sem sobreposição de informação. Neste contexto, a "Interface de Funções Administrativas" da Figura 3.2 tem grande importância, pois é responsável pelas associações feitas na cadeia e pela criação dos Indicadores. Apenas o administrador, sendo uma pessoa registada, tem acesso a esta interface.

O administrador pode criar, alterar ou apagar qualquer item desta cadeia (área, tema, grupo, indicador) e fazer as respectivas associações.

Na criação de um indicador, o administrador tem que indicar o Número de dimensões, Nome do indicador e a Fonte de dados. Para além destes, existem uns campos adicionais que indicam a forma de apresentação dos dados. O utilizador pode indicar se os dados se tratam de percentagem, reflectindo-se nas tabelas; se deseja observar os totais nas tabelas,

reflectindo-se pela soma vertical e horizontal dos dados e percentagens de cada valor em relação ao total; se pretende observar as percentagens nos gráficos.

O número de dimensões é extremamente importante para indicar à base de dados em que tabela gravar o indicador para futura representação. Como exemplo, criou-se o indicador "Abandono escolar no 1º ciclo do ensino básico" com três dimensões, tal como representado na figura 3.3, sendo esta a mensagem SQL gerada:

"INSERT INTO indicadores (iid,indicador, conceito, grupo, tipo,fonte,isperc,nosum,verp) VALUES (NULL,'Abandono escolar no 1º ciclo do ensino básico', 'Abandono escolar', 111,'factos_3',15,0,1,0)"

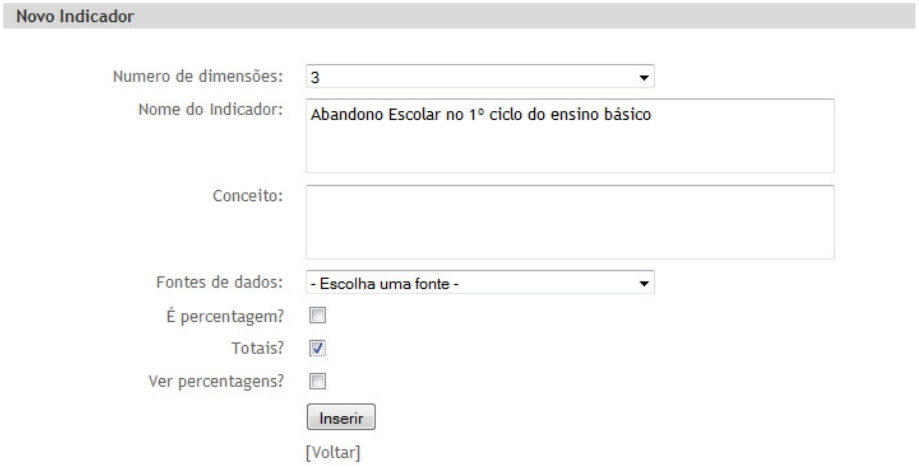


Figura 3.3: Exemplo da criação de indicador

Os parâmetros da mensagem variam de acordo com as escolhas feitas pelo utilizador. A construção dos indicadores tem de estar organizada por forma a garantir a correcta associação aos grupos. Uma vez criado o indicador é necessário atribuir nomes às dimensões e inserir valores para que se possam criar as relações.

Um indicador com, por exemplo, três dimensões, pode ser representado por:

$$I = f(D1,D2,D3)$$

onde D1 é a primeira dimensão, D2 a segunda e D3 a terceira. O indicador pode possuir um menor ou maior número de dimensões, tornando-se cada vez mais difícil de o representar visualmente quanto maior for esse número. De referir que a ordem de criação das dimensões será a ordem com que estas serão apresentadas na visualização. A ordem será descendente, ou seja, no caso da figura 3.4, o "Agrupamento Escolar" é a primeira dimensão, o "ano lectivo" a segunda e o "sexo" a terceira. Esta ordem reflecte-se na tabela tal como ilustrado na figura 3.5.

No caso do observatório em estudo, o indicador de três dimensões apresenta-se como a situação mais complexa, mas um dos objectivos deste trabalho é a implementação de indicadores com quatro dimensões. Ao registar as dimensões há que ter em consideração a sua ordem, uma vez que esta afectará a visualização do indicador.

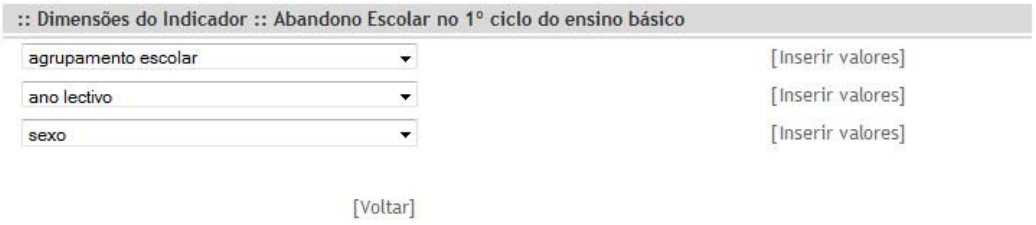


Figura 3.4: Exemplo da inserção das dimensões

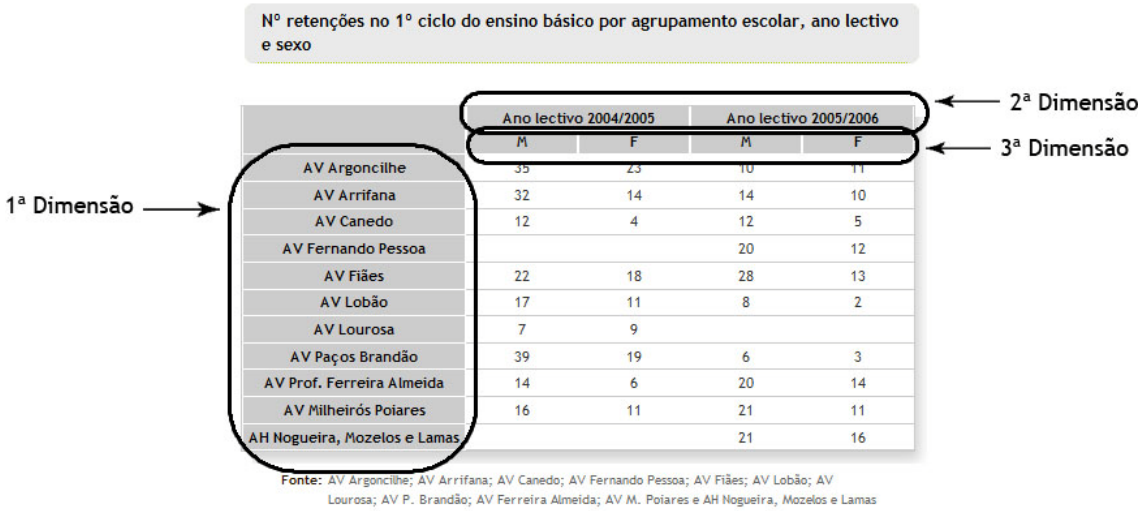


Figura 3.5: Representação da disposição das dimensões nas tabelas

Devido à complexidade do problema, nem sempre a melhor solução é implementada. Não tornando o sistema o mais simples possível, deparamo-nos muitas vezes com problemas na estrutura da base de dados.

Analisando a base de dados do Observatório foi possível verificar que ela é constituída por um elevado número de tabelas e vistas, que vão progressivamente aumentando. Centenas de tabelas e vistas que compõe a base de dados são originárias da criação de indicadores, pois sempre que é criado um indicador com X dimensões, são criados X tabelas que correspondem aos nomes dessas dimensões. Por exemplo, se se criar um indicador com as dimensões "Agrupamento Escolar" e "Ano Lectivo", vão ser automaticamente criadas duas tabelas com os respectivos nomes, concatenados com o ID do indicador (Agrupamento Escolar_ID), para suportar os registos dos valores das dimensões. Adicionalmente é criada uma vista que contêm os registos de todos os cruzamentos possíveis entre os valores dessas dimensões.

3.2 - Estudo dos Problemas

Segue-se um estudo detalhado dos problemas descritos na introdução desta tese.

3.2.1 - Indicadores de Quatro Dimensões

Um dos problemas encontrados no Observatório, é a presença de indicadores com a mesma informação para diferentes registos temporais, como são os casos da figura 3.6 e 3.7. Estes indicadores representam a mesma informação, ou seja, o "Insucesso no 1ºciclo do ensino básico por agrupamento escolar, nº alunos/retenções e ano escolar" mas para anos lectivos diferentes. Seria muito mais útil para o utilizador, se os indicadores fossem agrupados, isto é, se a informação proveniente dos dois indicadores em vista fosse unida de forma a constituir um só indicador com toda a informação. Para além de reduzir o número de indicadores necessários para representar a informação, melhora significativamente a visualização para o utilizador e permite um ganho de possibilidade de conhecimento, além de que pode permitir outras representações visuais, como por exemplo, observar o comportamento dos alunos em termos de retenções ao longo dos anos lectivos.

Insucesso escolar no 1º ciclo do ensino básico no ano lectivo 2004/2005 por agrupamento escolar, nº alunos/retenções e ano escolar

	Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	126	167	174	164	1	20	11	26
AV Arrifana	77	88	86	75	5	28	5	8
AV Canedo	91	116	95	91	0	12	0	4
AV Fernando Pessoa	269	294	282	236				
AV Fiães	119	141	142	149	0	22	5	13
AV Lobão	152	139	169	144	0	13	10	5
AV Lourosa					0	6	1	9
AV Paços Brandão	54	57	58	61	0	32	12	14
AV Prof. Ferreira Almeida	161	195	175	206	0	12	6	2
AV Milheirós Poiares	122	124	127	126	0	16	0	11
AH Nogueira, Mozelos e Lamas	205	207	234	217				

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

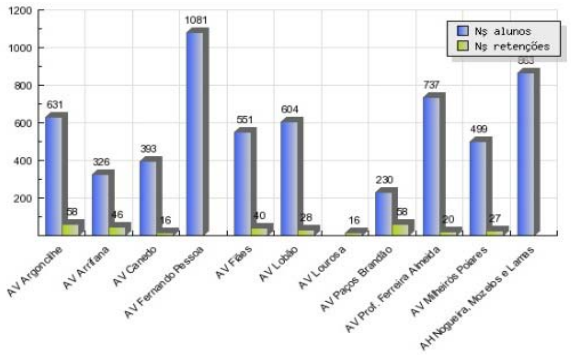


Figura 3.6: Insucesso escolar no 1º ciclo do ensino básico no ano lectivo 2004/2005 por agrupamento escolar, nº alunos/retenções e ano escolar

Insucesso escolar no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar, nº alunos/retenções e ano escolar

	Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	125	144	147	181	4	10	4	3
AV Arrifana	83	113	100	112	0	14	4	6
AV Canedo	93	105	96	94	0	8	4	5
AV Fernando Pessoa	226	262	280	277	0	17	3	12
AV Fiães	156	135	124	159	0	12	8	13
AV Lobão	133	156	138	153	0	6	0	4
AV Lourosa								
AV Paços Brandão	44	60	51	58	0	3	0	6
AV Prof. Ferreira Almeida	164	160	165	174	0	20	10	4
AV Milheirós Poiares	105	135	107	125	1	9	8	14
AH Nogueira, Mozelos e Lamas	191	218	213	248	0	21	7	9

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

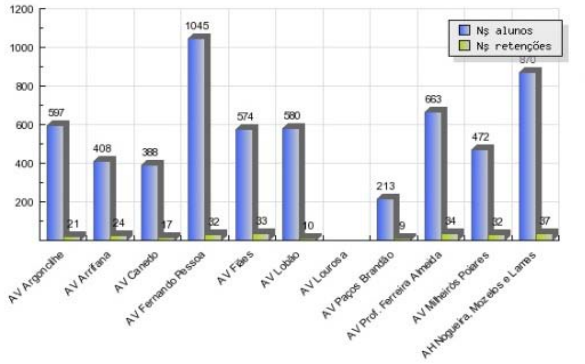


Figura 3.7: Insucesso escolar no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar, nº alunos/retenções e ano escolar

Devido a esta limitação, o utilizador foi também forçado a criar indicadores como o da figura 3.8, onde a informação é a mesma que se encontra nos indicadores das figuras 3.6 e 3.7, mas agora sem a dimensão "ano escolar" e com a dimensão "ano lectivo". Por exemplo, na figura 3.8, o número de retenções na AV Argoncilhe em 2004/2005 é a soma das retenções no 1º, 2º, 3º, e 4º ano da mesma escola na figura 3.7. Uma vez que há dados que são introduzidos mais que uma vez, existe o risco de incoerências devido a erro humano.

Uma vez que não é possível criar indicadores de quatro dimensões, os utilizadores foram forçados a criar inúmeros indicadores de outras dimensões de modo a poderem representar toda a informação

Deste modo a implementação de indicadores com quatro dimensões torna-se um dos requisitos do trabalho.

Insucesso escolar no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo e nº alunos/retenções

	Ano lectivo 2004/2005		Ano lectivo 2005/2006	
	Nº alunos	Nº retenções	Nº alunos	Nº retenções
AV Argoncilhe	631	58	597	21
AV Arrifana	326	46	408	24
AV Canedo	393	16	388	17
AV Fernando Pessoa	1081		1045	32
AV Fiães	551	40	574	33
AV Lobão	604	28	580	10
AV Lourosa		16		
AV Paços Brandão	230	58	213	9
AV Ferreira Almeida	737	20	663	34
AV Milheirós Poiares	499	27	472	32
AH Nogueira, Mozelos e Lamas	863		870	37

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

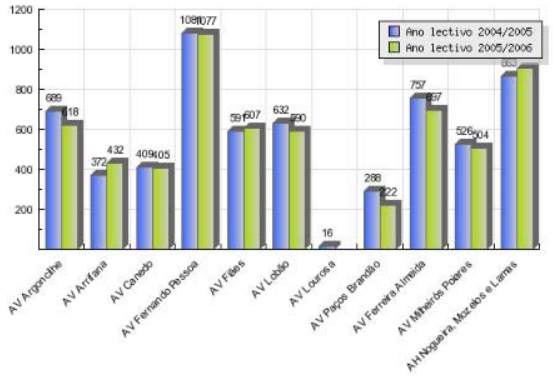


Figura 3.8: Exemplo de visualização de um indicador de três dimensões

3.2.2 - Nome dos Indicadores

A forma de construção do nome dos indicadores também constitui um problema. Actualmente é uma concatenação entre o nome inserido na sua criação pelo utilizador e o nome de cada uma das dimensões que o constituem.

No caso de indicadores simples, a solução funcionava bem. Contudo, em indicadores mais complexos ou que tentam contornar as limitações existentes há problemas.

No exemplo do indicador da figura 3.8, o nome é: "Insucesso escolar no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo e nº alunos/retenções".

Neste caso o nome introduzido pelo utilizador foi: "Insucesso escolar no 1º ciclo do ensino básico no ano lectivo 2005/2006" e as dimensões foram: "agrupamento escolar", "ano lectivo" e "nº alunos/retenções". Verifica-se pois que o nome do indicador não será o mais correcto, uma vez que na realidade representa o nº alunos e o nº retenções". Como o utilizador queria representar o nº de alunos e o nº de retenções num só indicador, teve necessidade de usar estes campos como uma dimensão. E uma vez que a descrição da dimensão também aparece no nome, não iria fazer sentido descrevê-lo como "Nº de Alunos e Nº de Retenções" em vez de "Insucesso escolar", pois aí o indicador iria ficar como "Nº de Alunos e Nº de Retenções no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar, ano lectivo e nº alunos/retenções".

A associação de vários indicadores leva a que o nome dos indicadores não fique correcto.

3.2.3 - Visualização

A visualização é um factor extremamente importante para a descoberta de conhecimento no Observatório. Na versão em estudo, a visualização suscita imensos problemas derivados das suas várias limitações.

Como é possível observar na figura 3.8, a representação gráfica para além de não ser a mais adequada, também não é, de todo, correcta. Os números representados nas colunas são a soma do número de alunos com o número de retenções. Na versão em estudo, qualquer que seja a situação analisada, a terceira dimensão é automaticamente somada de forma a poder ser representada graficamente. Isto para além de não ser correcto limita a percepção que os utilizadores possam ter dos dados, pois ao somar a terceira dimensão está, de certa forma, a eliminá-la. Para indicadores com uma ou duas dimensões o problema não ocorre, pois este tipo de representação permite visualizações até esse patamar. Assim sendo, esta representação não permite que os utilizadores observem os valores referentes à terceira dimensão de um modo distinto limitando-lhes, em muito, a quantidade de extracção de conhecimento possível.

Outra dificuldade que este tipo de gráficos encerra é a representação da dimensão espaço-temporal. Todos os dados presentes no Observatório possuem uma componente temporal, ou seja, foram registados numa determinada altura. Um dos requisitos das pessoas envolvidas no meio social é o de poderem detectar padrões e/ou tendências através da observação da evolução dos dados ao longo do tempo.

Como se pode observar na figura 3.8, a dimensão temporal está representada pela diferença de cores. Apesar de ter apenas dois valores (ano lectivo 2004/2005 e ano lectivo 2005/2006), este modo de representação não será o mais adequado para se vislumbrar tendências e ou padrões pois acumula-se muita informação num só gráfico, não dando muitas vezes ênfase ao pretendido. O caso complica-se se forem adicionados mais anos lectivos.

Nos exemplos das figura 3.9, 3.10, 3.11, que possuem respectivamente três, quatro e seis anos lectivos à medida que o número vai aumentando torna-se cada vez mais difícil para o utilizador perceber os gráficos. Na figura 3.10 para além da legenda estar mal colocada, os números tornam-se imperceptíveis e torna-se até complicado distinguir os diferentes anos.

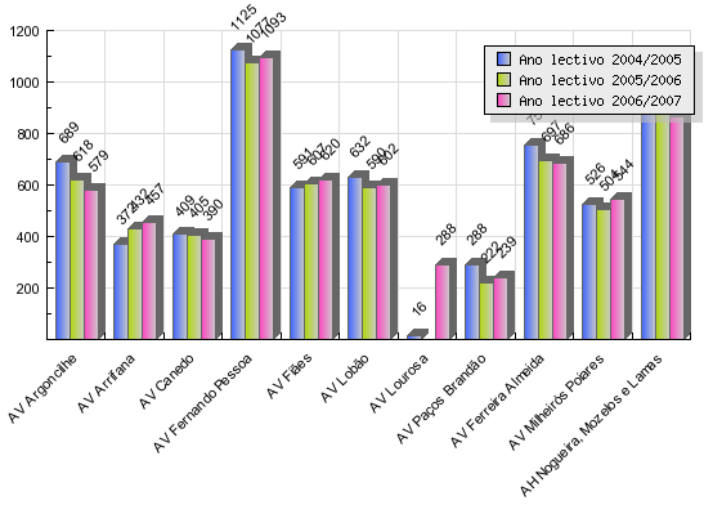


Figura 3.9: Exemplo do gráfico da figura 3.8, com três anos lectivos

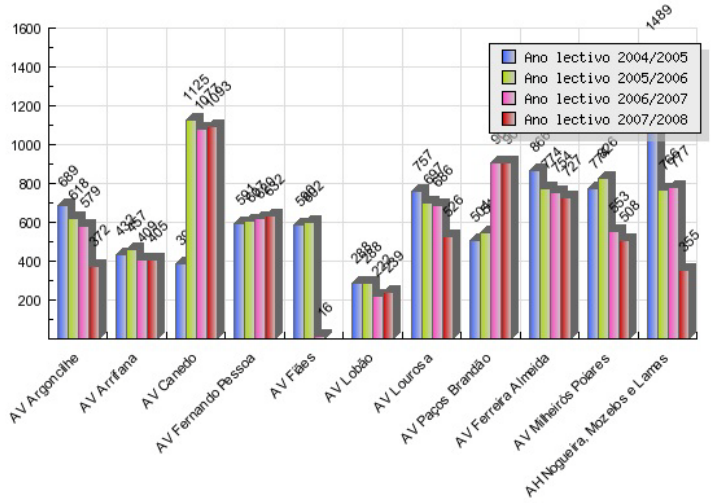


Figura 3.10:Exemplo do gráfico da figura 3.8, com quatro anos lectivos

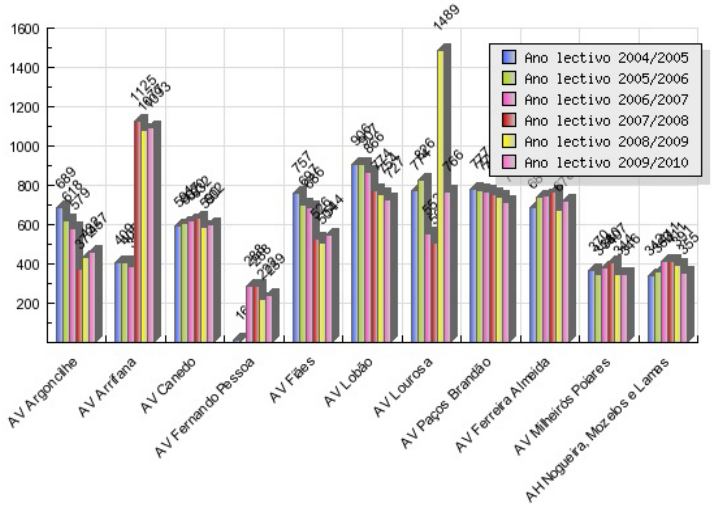


Figura 3.11:Exemplo do gráfico da figura 3.8, com seis anos lectivos

O facto de um indicador possuir uma grande quantidade de dados torna-se também um problema para a representação em tabelas HTML. Observando a tabela da figura 3.8, não se vislumbra qualquer inconsistência. Contudo, caso exista informação para mais anos lectivos, começam a surgir problemas de representação - figura 3.12, onde o indicador possui já seis anos lectivos. A tabela aumenta de tamanho, mas encolhe os dados.

A estrutura visual do Observatório fica igualmente deformado tal como exemplificado na figura 3.13. O utilizador terá assim que fazer um maior esforço para compreender a informação que pretende observar. O caso vai-se agravando à medida que o número de valores da segunda ou e/ou terceira dimensão aumenta.

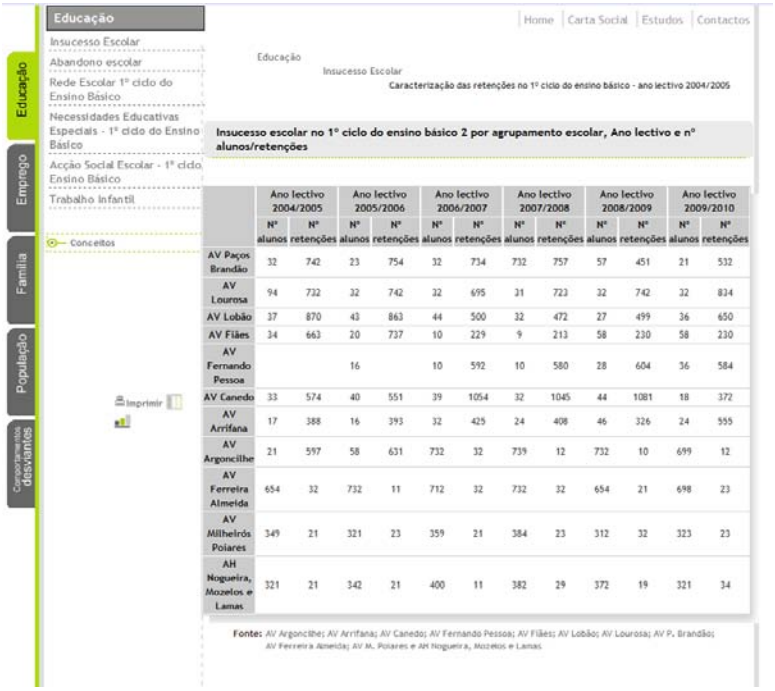


Figura 3.12: Exemplo da tabela de um indicador com seis anos lectivos

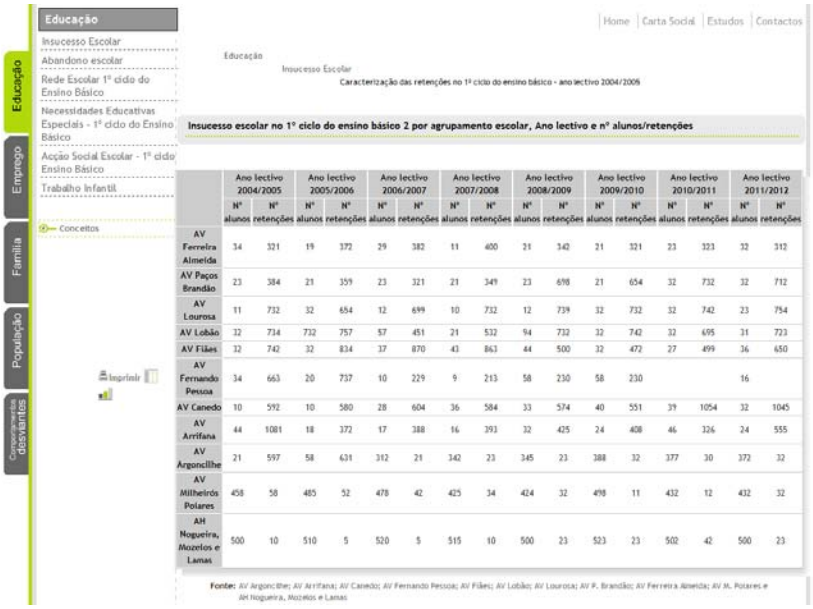


Figura 3.13: Exemplo da tabela de um indicador com oito anos lectivos

Outro dos problemas da visualização de dados espaço-temporais está bem presente nos indicadores que possuem a dimensão freguesias. A visualização é feita aqui através de mapas, como o da figura 3.14, e tal como no caso anterior, os administradores são obrigados a criar vários indicadores para as diferentes dimensões temporais. Isto acontece por não haver forma de demonstrar no mesmo indicador os diferentes mapas para os diferentes registos temporais. A criação de indicadores para cada registo provoca um grande aumento no número de tabelas, dificultando a sua procura e acesso.

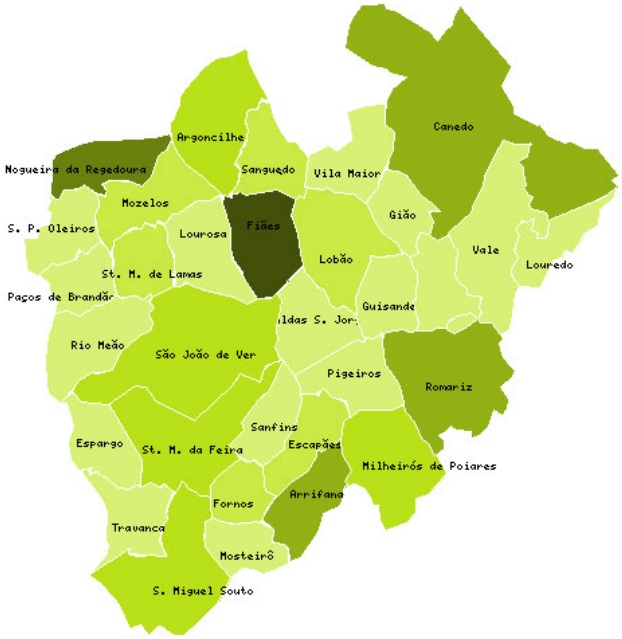


Figura 3.14: Exemplo de visualização em mapas

3.2.4 - Padrões de Utilização

Estando o Observatório Social em constante evolução e crescimento, é do interesse das pessoas nele envolvidas oferecer aos utilizadores os conteúdos que mais lhes possam interessar. Por outro lado, como se trata de um ambiente social, pode ser importante dar enfoque às várias questões sociais, analisando as áreas ou indicadores que mais afectam e preocupam os cidadãos. Qualquer que seja a finalidade será sempre necessário implementar uma plataforma que permita extrair informação relevante sobre os utilizadores do Observatório e os seus hábitos, como páginas mais visualizadas, número, percursos, entre outros. Pretende-se extrair essa informação e apresenta-la aos utilizadores.

Na versão inicial não havia qualquer gestão de padrões de utilização.

3.2.5 - Gestão de Utilizadores do Painel de Administração

Na versão inicial do *BackOffice* do Observatório Social de Santa Maria da Feira, existe apenas um tipo de utilizador, o Administrador. Este é o responsável por toda a gestão do Observatório, isto é, é o responsável por criar, alterar ou apagar áreas, temas, fontes, tipos de dimensão, grupos e indicadores. É também o responsável pelos dados inseridos nos indicadores.

A dificuldade de haver só um tipo de utilizador responsável pela gestão do Observatório, é que nem sempre as pessoas que inserem dados estão qualificadas para criar indicadores, pois estes seguem um determinado tipo de regras que não são de conhecimento geral. Inserindo novos perfis de utilizador poder-se-á alargar a utilização e interactividade do Observatório, dando a possibilidade de tornar o processo de inserção de dados mais abrangente, não se resumindo apenas ao administrador.

3.3 - Os Primeiros Problemas e as Primeiras Opções

O Observatório Social é um projecto que se encontra actualmente alojado nos servidores do IDIT. De forma a dar início ao trabalho, foi necessário criar uma cópia do servidor num computador pessoal, copiando todos ficheiros do Observatório.

No decorrer do processo, o primeiro problema surgiu aquando da extracção da base de dados. Na extracção desta os problemas ocorriam sempre em tabelas e vistas que advinham da criação de indicadores. Analisando a origem desta situação, foi possível detectar que existiam inúmeras tabelas que não possuíam quaisquer dados, e vistas que continham erros. Optou-se por extrair apenas as tabelas essenciais ao funcionamento do Observatório, ou seja, livre de quaisquer indicadores.

Para a instalação do servidor foi usado o *WampServer* em ambiente Windows. Este usa o Apache 2.2.11, PHP 5.3.0 e MYSQL 5.1.36.

3.3.1 - Versões Antigas

Grande parte das versões de programas, linguagens de programação, bibliotecas, etc., vai sofrendo actualizações ao longo do tempo. O PHP não foge à regra. Tendo sido o Observatório criado com a versão 4 do PHP e estando este actualmente na versão 5, resolveu-se fazer a devida actualização, pois achou-se que não seria boa regra estar a usar versões antigas na dissertação. Várias funções de PHP usadas no Observatório tiveram que ser substituídas por funções idênticas, uma vez que se tornaram obsoletas e já não são suportadas.

A biblioteca *JpGraph*, responsável por criar os gráficos no Observatório foi também actualizada para a sua versão mais recente, 3.0.7, de forma a funcionar com o PHP5, apesar de até então não estar ainda decidida a sua utilização.

3.3.2 - Melhorias no Observatório

Com uma versão pronta para poder ser trabalhada, houve, inicialmente, a necessidade de analisar as funcionalidades do Observatório e de perceber o seu funcionamento. Um programador, ao pegar num sistema implementado, que contêm várias dezenas de ficheiros PHP, necessita de passar por uma fase de reconhecimento do código, da sua estrutura, e da funcionalidade de todos os ficheiros ou de pelo menos da parte indispensável ao correcto funcionamento da aplicação.

Antes de avançar para o processo de verificação de todas as funcionalidades do Observatório, tornou-se indispensável compreender o motivo da existência de tantas tabelas sem dados e de vistas com erros, para que o futuro trabalho no Observatório não fosse afectado. Como estas tabelas e vistas surgiam após a criação dos indicadores, depreendeu-se que o problema estaria relacionado com algum dos passos que culminavam nessa criação. A origem do problema foi descoberta após se ter criado e, posteriormente, eliminado um indicador. A acção de eliminação do indicador apenas eliminava uma das tabelas da base de dados referenciadas pelo indicador em vez de eliminar todas as tabelas e/ou vistas que tinham sido geradas no seu processo de criação. A base de dados ficava a acumular tabelas que já não deveriam existir, e as vistas, que também já não deveriam existir, declaravam erros por não encontrarem as referências dos indicadores correspondentes. Esta situação originava incongruência de dados e impedia a total extracção da informação da base de dados. Deste modo, tornou-se indispensável corrigir o processo, por forma a permitir que o

utilizador, sempre que eliminasse um indicador, eliminasse também todas as tabelas e vistas devidas.

Enquanto se estudava o problema anterior, constatou-se uma outra situação na página de "inserção de dados" dos indicadores. Se um utilizador acedesse à página de "inserção de dados" dos indicadores e saísse desta sem inserir qualquer informação, o indicador ficava automaticamente bloqueado e tornava-se impossível inserir novos dados. Foram, por isso, feitas alterações no código PHP para corrigir esta situação.

Após a correcção destas adversidades, testaram-se todas as funcionalidades do Observatório por forma a melhor o entender. Com isto, foram descobertos mais alguns erros que, tal como nos casos anteriores, urgiu corrigir. Derivado da complexidade de algumas situações, este processo ainda demorou bastante tempo. De seguida são apresentados os bugs, erros, problemas que necessitaram de correcção:

1. Depois de instalada uma versão do servidor em ambiente Windows, a biblioteca *JpGraph* deixou de funcionar correctamente por falta de permissões. Isto aconteceu porque sempre que um gráfico era alterado, o *JpGraph* tentava apagar o respectivo ficheiro de imagem em cache já existente e não conseguia por falta de permissões. Para resolver este problema foi necessário implementar uma função que permitisse apagar os ficheiros directamente via PHP.
2. Problemas na página de "alteração de indicadores". Um dos erros ocorria na mudança do nome do indicador. Caso o utilizador alterasse o nome, tornava-se impossível editar os dados. Isto acontecia porque a mudança do nome não era reflectida em todas as tabelas e/ou vistas da base de dados onde esse nome se encontrava representado.
3. Um bug extremamente relevante surgiu na página de "inserção dos dados" dos indicadores. Quando o administrador acrescentava um valor numa dimensão, os dados correspondentes ao cruzamento das dimensões não coincidiam. De modo a exemplificar e a perceber-se melhor o problema, foi criado um indicador com três dimensões: "agrupamento escolar", "ano escolar" e "nº alunos/retenções". Para cada dimensão foram inseridos vários valores. De seguida foram inseridos vários dados nos cruzamentos dos valores das dimensões. Depois de inseridos os dados, voltando à mesma página é possível verificar que estes são apresentados correctamente, isto é, cada valor apresentado corresponde ao cruzamento dos valores das diferentes dimensões correcto tal como está representado na figura 3.15. Contudo, caso o utilizador acrescentasse, por exemplo, um ano lectivo, os dados surgiam incorrectamente tal como está na figura 3.16. Neste caso adicionou-se o ano lectivo 2005/2006 e, como é possível observar, os dados são colocados de forma contínua, não respeitando os cruzamentos dos valores das diferentes dimensões. Esta situação ocorria pelo facto de não existir uma relação na forma como os dados e os cruzamentos das dimensões são apresentados, isto é, o modo como o cruzamento das dimensões é apresentada, é independente do modo que os dados destas relações são apresentados. Foi essencial corrigir este bug para que os utilizadores responsáveis pela inserção de dados não necessitassem de inserir os dados todos novamente sempre que acrescentassem um novo valor a uma dimensão.

Insucesso Escolar no 1º ciclo do ensino básico			
agrupamento escolar	ano lectivo	nº alunos/retenções	Valor
AV Argoncilhe	Ano lectivo 2004/2005	Nº alunos	631.0
AV Argoncilhe	Ano lectivo 2004/2005	Nº retenções	58.0
AV Arrifana	Ano lectivo 2004/2005	Nº alunos	326.0
AV Arrifana	Ano lectivo 2004/2005	Nº retenções	46.0
AV Canedo	Ano lectivo 2004/2005	Nº alunos	393.0
AV Canedo	Ano lectivo 2004/2005	Nº retenções	16.0
AV Fernando Pessoa	Ano lectivo 2004/2005	Nº alunos	1081.0
AV Fernando Pessoa	Ano lectivo 2004/2005	Nº retenções	-1.0
AV Fiães	Ano lectivo 2004/2005	Nº alunos	551.0
AV Fiães	Ano lectivo 2004/2005	Nº retenções	40.0
AV Lobão	Ano lectivo 2004/2005	Nº alunos	604.0
AV Lobão	Ano lectivo 2004/2005	Nº retenções	28.0
AV Lourosa	Ano lectivo 2004/2005	Nº alunos	-1.0
AV Lourosa	Ano lectivo 2004/2005	Nº retenções	16.0
AV Paços Brandão	Ano lectivo 2004/2005	Nº alunos	230.0
AV Paços Brandão	Ano lectivo 2004/2005	Nº retenções	58.0
AV Ferreira Almeida	Ano lectivo 2004/2005	Nº alunos	737.0
AV Ferreira Almeida	Ano lectivo 2004/2005	Nº retenções	20.0
AV Milheirós Poiares	Ano lectivo 2004/2005	Nº alunos	499.0
AV Milheirós Poiares	Ano lectivo 2004/2005	Nº retenções	27.0
AH Nogueira, Mozelos e Lamas	Ano lectivo 2004/2005	Nº alunos	863.0
AH Nogueira, Mozelos e Lamas	Ano lectivo 2004/2005	Nº retenções	-1.0
Inserir			

Figura 3.15: Página “Inserir dados”

Insucesso Escolar no 1º ciclo do ensino básico

agrupamento escolar	ano lectivo	nº alunos/retenções	Valor
AV Argoncilhe	Ano lectivo 2004/2005	Nº alunos	631.0
AV Argoncilhe	Ano lectivo 2005/2006	Nº alunos	58.0
AV Argoncilhe	Ano lectivo 2004/2005	Nº retenções	326.0
AV Argoncilhe	Ano lectivo 2005/2006	Nº retenções	46.0
AV Arrifana	Ano lectivo 2004/2005	Nº alunos	393.0
AV Arrifana	Ano lectivo 2005/2006	Nº alunos	16.0
AV Arrifana	Ano lectivo 2004/2005	Nº retenções	1081.0
AV Arrifana	Ano lectivo 2005/2006	Nº retenções	-1.0
AV Canedo	Ano lectivo 2004/2005	Nº alunos	551.0
AV Canedo	Ano lectivo 2005/2006	Nº alunos	40.0
AV Canedo	Ano lectivo 2004/2005	Nº retenções	604.0
AV Canedo	Ano lectivo 2005/2006	Nº retenções	28.0
AV Fernando Pessoa	Ano lectivo 2004/2005	Nº alunos	-1.0
AV Fernando Pessoa	Ano lectivo 2005/2006	Nº alunos	16.0
AV Fernando Pessoa	Ano lectivo 2004/2005	Nº retenções	230.0
AV Fernando Pessoa	Ano lectivo 2005/2006	Nº retenções	58.0
AV Fiães	Ano lectivo 2004/2005	Nº alunos	737.0
AV Fiães	Ano lectivo 2005/2006	Nº alunos	20.0
AV Fiães	Ano lectivo 2004/2005	Nº retenções	499.0
AV Fiães	Ano lectivo 2005/2006	Nº retenções	27.0
AV Lobão	Ano lectivo 2004/2005	Nº alunos	863.0
AV Lobão	Ano lectivo 2005/2006	Nº alunos	-1.0
AV Lobão	Ano lectivo 2004/2005	Nº retenções	
AV Lobão	Ano lectivo 2005/2006	Nº retenções	
AV Lourosa	Ano lectivo 2004/2005	Nº alunos	
AV Lourosa	Ano lectivo 2005/2006	Nº alunos	
AV Lourosa	Ano lectivo 2004/2005	Nº retenções	
AV Lourosa	Ano lectivo 2005/2006	Nº retenções	
AV Paços Brandão	Ano lectivo 2004/2005	Nº alunos	
AV Paços Brandão	Ano lectivo 2005/2006	Nº alunos	
AV Paços Brandão	Ano lectivo 2004/2005	Nº retenções	
AV Paços Brandão	Ano lectivo 2005/2006	Nº retenções	
AV Ferreira Almeida	Ano lectivo 2004/2005	Nº alunos	
AV Ferreira Almeida	Ano lectivo 2005/2006	Nº alunos	
AV Ferreira Almeida	Ano lectivo 2004/2005	Nº retenções	
AV Ferreira Almeida	Ano lectivo 2005/2006	Nº retenções	
AV Milheirós Poiares	Ano lectivo 2004/2005	Nº alunos	
AV Milheirós Poiares	Ano lectivo 2005/2006	Nº alunos	
AV Milheirós Poiares	Ano lectivo 2004/2005	Nº retenções	
AV Milheirós Poiares	Ano lectivo 2005/2006	Nº retenções	
AH Nogueira, Mozelos e Lamas	Ano lectivo 2004/2005	Nº alunos	
AH Nogueira, Mozelos e Lamas	Ano lectivo 2005/2006	Nº alunos	
AH Nogueira, Mozelos e Lamas	Ano lectivo 2004/2005	Nº retenções	
AH Nogueira, Mozelos e Lamas	Ano lectivo 2005/2006	Nº retenções	

Inserir

Figura 3.16: Página "inserir dados" após acrescentar o ano lectivo 2005/2006

4. Se um indicador possuir a dimensão "freguesias", a visualização gráfica é feita através de mapas. Um ponto que se achou relevante alterar, foi a estrutura da base de dados em relação à dimensão "freguesias". Quando se inseria esta dimensão num indicador, era criava uma vista para o respectivo indicador, com os dados da tabela "freguesias". O grande problema da criação destas vistas, é que se o utilizador alterasse ou eliminasse uma freguesia, isso iria afectar também a tabela freguesias o que "destruía" a base de dados em relação a esta dimensão e deixava de ser possível implementar futuros indicadores com a dimensão. Como tal, em vez de vistas foram adoptadas tabelas onde o utilizador poderá alterar ou apagar sem qualquer risco de afectar a construção da base de dados e de futuros indicadores.
5. Ao eliminar um valor numa dimensão que já possuísse dados, o Observatório apenas apagava a sua referência na tabela da dimensão respectiva. Contudo os dados nas restantes tabelas e vistas que foram criados relativos a essa dimensão apagada permaneciam inalteráveis. Se o utilizador resolvesse inserir novamente o valor daquela dimensão que tinha anteriormente apagado, o Observatório não mostrava os dados correctamente. Isto acontecia devido ao ID do valor recriado e o ID que os dados possuíam deste valor serem diferentes. Foi necessário corrigir este problema, de modo a garantir que se um utilizador apagar um valor duma dimensão com dados, os dados referentes a essa dimensão em todas as tabelas e vistas serão igualmente apagados.

Para resolver todas estas situações, foi necessário efectuar uma significativa reestruturação da base de dados. Tendo solucionado a maioria dos problemas que poderiam afectar o desenvolvimento desta tese, e tendo entendido de uma forma abrangente o modo de funcionamento do Observatório, foi possível dar início ao cumprimento dos objectivos primordiais desta dissertação.

3.4 - Percurso, Propostas de Soluções e as Tecnologias

Tendo em conta o que foi dito no estudo do problema, o percurso a efectuar pode ser resumido em cinco passos:

1. Implementação de indicadores com quatro dimensões;
2. Implementação da solução para o problema do nome dos indicadores;
3. Implementação das soluções para os problemas da Visualização;
4. Implementação de gestão de utilizadores no painel de administração do Observatório;
5. Implementação de registo de padrões de utilização (*Web Usage Mining)*;

Para o cumprimento dos objectivos desta tese, optou-se por continuar a usar as linguagens PHP, HTML, JavaScript e SQL, visto que, comprovadamente, continuam a ser actuais.

3.4.1 - Indicadores de Quatro Dimensões

Decidiu-se avançar no imediato para a implementação de indicadores com quatro dimensões, uma vez que este processo é parte da solução para com os problemas detectados.

O único aspecto a referir é que a sua implementação segue os mesmos moldes de criação dos indicadores com menor número de dimensões e como tal, foi necessário um estudo pormenorizado de como esta criação era efectuada.

3.4.2 - Nome dos Indicadores

Como este problema se devia à associação de vários indicadores, a primeira alternativa considerada para o resolver seria a implementação de uma interface que permitisse que o utilizador, ao criar os indicadores separadamente, pudesse dar a indicação que os queria agregar, ou seja mostrar os vários indicadores num só. Assim, se o utilizador criasse, por exemplo, os indicadores "Nº alunos no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar e ano escolar" e "Nº retenções no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar e ano escolar", e indicasse que os queria agregar num só, este ficaria "Nº alunos e Nº retenções no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar e ano escolar". O problema desta solução é que iria tornar o processo de construção de indicadores muito mais complicado, exigindo mais trabalho ao utilizador. O nome dos indicadores teria que ser processado para formar um único e as tabelas teriam que ser agregadas, sendo criada automaticamente uma nova dimensão para separar e indicar na tabela a que parâmetros pertenciam os dados. Isto iria alterar grande parte da estrutura de criação de indicadores do Observatório.

Pensando numa alternativa mais simples, a proposta de solução apresentada incide sobre o nome das dimensões. Se o problema está em aparecer a dimensão X no nome do indicador, então há que dar a possibilidade ao utilizador de retirar a dimensão X do nome do indicador. Deste modo, o utilizador poderia criar o indicador com a designação que bem entendesse. No caso do exemplo acima, excluindo do nome a terceira dimensão, "nº alunos/retenções", ficaria "Nº alunos e retenções no 1º ciclo do ensino básico no ano lectivo 2005/2006 por agrupamento escolar e ano escolar".

3.4.3 - Visualização

O próximo passo seria o da resolução do problema de visualização. Devido à necessidade de estudar todos os casos possíveis de cruzamento de dados e devido à escassez de soluções existentes, este problema consumiu bastante tempo. O primeiro ponto de reflexão prendeu-se com a visualização em tabelas. Com a implementação de indicadores com quatro dimensões, este problema agravava-se. Tornava-se necessário implementar uma solução que permitisse observar os dados num indicador, por maior que a sua quantidade fosse, de uma forma simples e eficaz, por mais reduzido que o espaço fosse.

Em todos os indicadores do Observatório a dimensão temporal encontra-se presente, seja directamente, quando entra efectivamente nas dimensões do indicador, ou indirectamente, quando entra apenas no nome do indicador, devido ao facto de o Observatório não funcionar com indicadores de quatro dimensões e, deste modo, os utilizadores serem forçados a apenas referir essa dimensão no nome quando tinham mais três dimensões a representar. Sendo possível criar indicadores com quatro dimensões os utilizadores poderiam colocar a dimensão temporal directamente em todos os indicadores.

A proposta de solução consiste em dividir os dados pela dimensão temporal, em tabelas. No exemplo da figura 3.17, se a tabela for dividida em tabelas de dimensão inferior pela dimensão temporal ficaria como está representado na figura 3.18. O problema estaria na

forma de fornecer ao utilizador a possibilidade de escolher a tabela que desejaria observar, não estando estas todas presentes em simultâneo. A solução teria que funcionar para todos os indicadores, quer estes possuíssem duas, três ou mesmo quatro dimensões. Seria também necessário que não tornasse a visualização lenta - factor extremamente importante. Foram estudadas várias hipóteses, avaliando qual seria a mais eficaz, útil e de fácil manuseamento, de forma a simplificar o acesso por parte do utilizador, e de modo a que este possa tirar o maior proveito possível da visualização das tabelas.

	Ano lectivo 2004/2005		Ano lectivo 2005/2006		Ano lectivo 2006/2007		Ano lectivo 2007/2008	
	Nº alunos	Nº retenções	Nº alunos	Nº retenções	Nº alunos	Nº retenções	Nº alunos	Nº retenções
AV Argoncilhe	631	58	597	21	600	50	600	50
AV Arrifana	326	46	408	24	400	30	400	30
AV Canedo	393	16	388	17	400	20	400	20
AV Fernando Pessoa	1081		1045	32	1000	40	1000	50
AV Fiães	551	40	574	33	600	50	600	50
AV Lobão	604	28	580	10	600	30	600	30
AV Lourosa		16						
AV Paços Brandão	230	58	213	9	220	40	230	11
AV Ferreira Almeida	737	20	663	34	700	30	700	30
AV Milheirós Poiares	499	27	472	32	500	30	500	30
AH Nogueira, Mozelos e Lamas	863		870	37	900	40	900	40

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

Figura 3.17: Nº de alunos e Retenções no 1º ciclo do ensino básico por agrupamento escolar e ano lectivo

	Ano lectivo 2004/2005	
	Nº alunos	Nº retenções
AV Argoncilhe	631	58
AV Arrifana	326	46
AV Canedo	393	16
AV Fernando Pessoa	1081	
AV Fiães	551	40
AV Lobão	604	28
AV Lourosa		16
AV Paços Brandão	230	58
AV Ferreira Almeida	737	20
AV Milheirós Poiares	499	27
AH Nogueira, Mozelos e Lamas	863	

	Ano lectivo 2005/2006	
	Nº alunos	Nº retenções
AV Argoncilhe	597	21
AV Arrifana	408	24
AV Canedo	388	17
AV Fernando Pessoa	1045	32
AV Fiães	574	33
AV Lobão	580	10
AV Lourosa		
AV Paços Brandão	213	9
AV Ferreira Almeida	663	34
AV Milheirós Poiares	472	32
AH Nogueira, Mozelos e Lamas	870	37

	Ano lectivo 2006/2007	
	Nº alunos	Nº retenções
AV Argoncilhe	600	50
AV Arrifana	400	30
AV Canedo	400	20
AV Fernando Pessoa	1000	40
AV Fiães	600	50
AV Lobão	600	30
AV Lourosa		
AV Paços Brandão	220	40
AV Ferreira Almeida	700	30
AV Milheirós Poiares	500	30
AH Nogueira, Mozelos e Lamas	900	40

	Ano lectivo 2005/2006	
	Nº alunos	Nº retenções
AV Argoncilhe	597	21
AV Arrifana	408	24
AV Canedo	388	17
AV Fernando Pessoa	1045	32
AV Fiães	574	33
AV Lobão	580	10
AV Lourosa		
AV Paços Brandão	213	9
AV Ferreira Almeida	663	34
AV Milheirós Poiares	472	32
AH Nogueira, Mozelos e Lamas	870	37

Figura 3.18: Nº de alunos e Retenções no 1º ciclo do ensino básico por agrupamento escolar e ano lectivo dividido em quatro tabelas pela segunda Dimensão.

A primeira solução reflectida é a mais comum - usar uma *drop down box*, figura 3.19. Para cada caso, a *drop down box* iria conter os valores da dimensão em causa, e através desta, o

utilizador poderia escolher a tabela que desejaria observar, bastando clicar no valor da dimensão que pretendesse. Contudo, esta solução não seria a mais eficiente nem a de mais fácil utilização, pois, sempre que o utilizador quisesse observar outros valores, teria que abrir a *drop down box* e navegar por esta à procura do valor que pretendesse. Quando escolhesse o valor da dimensão, essa escolha teria de ser enviada para o servidor de forma a carregar a página com os valores referentes à escolha. A navegação não iria ser a melhor, podendo tornar a visualização um processo ainda mais demorado. Como tal, foi pensado um método que não exigisse o recarregamento das páginas para navegar nos diferentes valores da dimensão, e que fosse mais eficaz e simples para os utilizadores procurarem o respectivo valor da dimensão que desejassem observar.

O proposto é a implementação de separadores através de JavaScript e AJAX, figura 3.20, para representar os diferentes valores da segunda dimensão que, através de AJAX, recolhem os conteúdos dinamicamente sem a necessidade de recarregar a página, tornando a visualização muito mais simples, eficaz e interactiva.

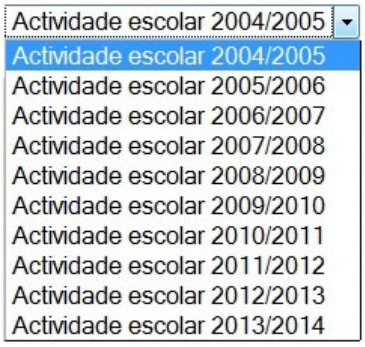


Figura 3.19: Exemplo de uma Dropbox

Figura 3.20: Exemplo de quatro separadores

- **AJAX** – Conjunto de técnicas de desenvolvimento *Web* usadas para aumentar a interactividade com o cliente. Usando AJAX, as aplicações *Web* podem enviar dados do servidor para o cliente sem ter que recarregar a página inteira. O AJAX pode também ser usado para o servidor recolher dados da aplicação *Web* de um

modo transparente. Em [33] o autor escreve sobre as tecnologias que o AJAX incorpora assim com as suas funcionalidades.

Em relação à visualização gráfica, esta estava sujeita a vários requisitos que elevavam o grau de complexidade:

1. A representação gráfica tinha que permitir visualizar uma, duas, três e quatro dimensões.
2. Devia ter em conta que são dados espaço-temporais e, como tal, não pode descurar a importância destas dimensões.
3. A visualização necessitava de estar dirigida a pessoas instruídas que pretendessem retirar algum tipo de conhecimento como padrões e tendências, e dirigida a pessoas comuns, sem grande instrução na área, que apenas quisessem observar os dados de uma forma gráfica e que os compreendessem.

Analisando o que foi dito no capítulo 2 sobre esta questão e pesquisando por várias aplicações capazes de criar visualizações gráficas, foi possível constatar que não existe propriamente uma forma visual simples de representar dados de três e quatro dimensões. Muitas das formas de representação encontradas seriam extremamente complicadas para grande parte dos utilizadores e limitadas aos seus casos de utilização.

Outro facto limitador importante é que não era possível escolher um software ou aplicação de visualizações gráficas que não fosse possível de enquadrar no Observatório, ou seja, no ambiente *Web*. Para resolver este problema de visualização gráfica, foi necessário colocarmo-nos no lado do utilizador, de forma a percebermos o que deseja este observar nos dados, e como os pretende observar.

Segundo *Edward Tufte* [25], todas as visualizações devem, para além de exibir os dados, induzir o utilizador a pensar na substância em vez da metodologia usada, do design, ou outros factores. Devem evitar distorcer o que os dados têm para dizer. Devem servir para apresentar elevadas quantidades de números num espaço pequeno; tornar grandes conjuntos de dados coerentes; revelar os dados em vários níveis de detalhe.

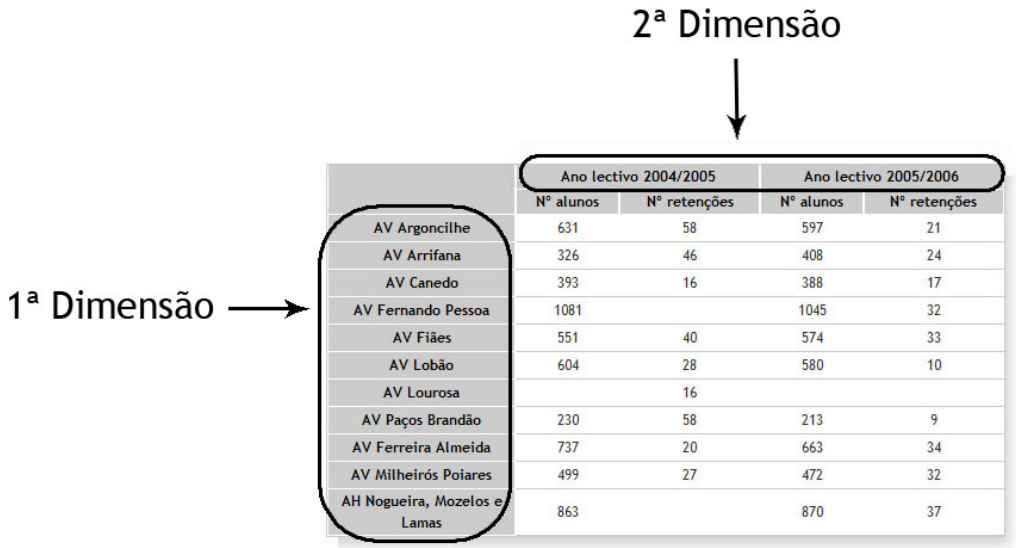
No caso do Observatório, os dados apresentados pelos indicadores, variam todos no tempo, seja por ano, ano lectivo, censos, etc. Não sendo possível implementar visualizações complexas, resolveu-se adoptar métodos de visualização simples. Os gráficos actualmente presentes no Observatório são gráficos de colunas, que acabam por ser os mais adequados e mais simples para o comum utilizador, contudo, este tipo de gráficos não permite nos casos mais complexos visualizar todas as dimensões pretendidas.

A proposta de solução, passa por dar a possibilidade ao utilizador para, dentro do indicador, escolher o que deseja observar. Isto é possível obter através de uma tabela dinâmica onde o utilizador pode escolher de que modo quer observar a evolução dos dados no gráfico. Usando a primeira e a segunda dimensão, os utilizadores, clicando nos diferentes valores destas, podem observar diferentes gráficos. Se clicar num dos valores da primeira dimensão podem observar como esse valor se comporta ao longo dos valores da segunda dimensão e vice-versa. Por exemplo, observando a figura 3.21, se o utilizador pretender observar o nº alunos e o nº retenções para a escola de "AV Argoncilhe" ao longo dos diferentes anos lectivos basta clicar nesse valor, sendo apresentado o respectivo gráfico. Caso pretenda observar o que aconteceu para o "ano lectivo 2005/2006" em todas as escolas, basta também clicar nesse valor.

Com esta solução já se torna possível representar as três dimensões de um indicador. Contudo, a utilização de gráficos normal de colunas com esta solução não permite

representar a quarta dimensão. Para esta, a solução é usar gráficos de colunas empilhadas tal como está representado na figura 3.22.

Para a representação dos gráficos foi escolhida, tal como já vinha sendo usada, a biblioteca PHP, *JpGraph*. Para criar uma interacção limpa e rápida a escolha recaiu sobre o JavaScript e o AJAX.



	Ano lectivo 2004/2005		Ano lectivo 2005/2006	
	Nº alunos	Nº retenções	Nº alunos	Nº retenções
AV Argoncilhe	631	58	597	21
AV Arrifana	326	46	408	24
AV Canedo	393	16	388	17
AV Fernando Pessoa	1081		1045	32
AV Fiães	551	40	574	33
AV Lobão	604	28	580	10
AV Lourosa		16		
AV Paços Brandão	230	58	213	9
AV Ferreira Almeida	737	20	663	34
AV Milheirós Poiares	499	27	472	32
AH Nogueira, Mozelos e Lamas	863		870	37

Figura 3.21: 1ª e 2ª Dimensão

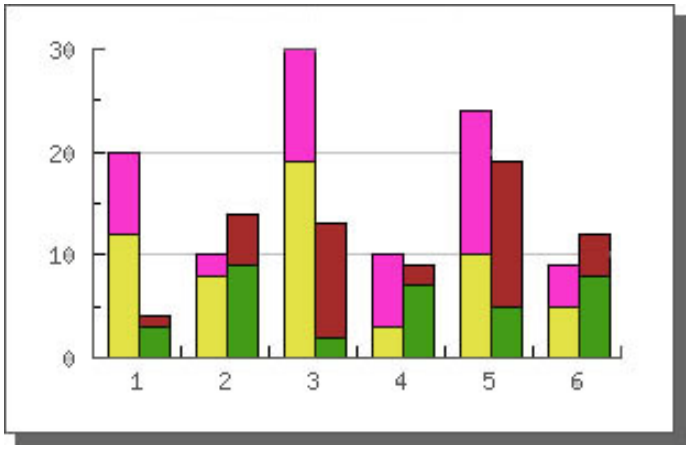


Figura 3.22: Gráfico de Colunas Empilhadas

A solução das tabelas interactivas é também a solução proposta para o problema da visualização em mapas, de forma a demonstrar no mesmo indicador os diferentes mapas para os diferentes registos temporais. O utilizador ao clicar nos diferentes registos temporais, será lhe apresentado os diferentes mapas correspondentes a esses registos.

3.4.4 - Gestão de Utilizadores

Após o desenvolvimento destas soluções dos problemas de visualização, que foi um processo bastante demorado devido sobretudo às várias dimensões, as atenções viraram-se

para a gestão de utilizadores do painel de administração. Antes de se ter implementado esta funcionalidade foi necessário efectuar um estudo sobre os utilizadores deste painel e do processo que leva à criação dos indicadores.

A proposta baseia-se em três tipos de utilizadores: administrador, gestor e técnico, todos com diferentes permissões e cargos. O administrador, não possui qualquer restrição, sendo este o responsável por criar e modificar utilizadores. O gestor possui todas as permissões do administrador excepto as acções envolvendo utilizadores. A ideia do gestor é que seja atribuído às pessoas instruídas para criar, alterar ou apagar indicadores, temas, áreas, etc. O técnico apenas pode inserir ou alterar dados dos indicadores. Esta ideia surge sobretudo devido aos problemas anexados à criação dos indicadores porque nem sempre as pessoas que inserem ou possuem os dados estão aptas para os criar.

Pode assim facilitar a gestão do Observatório de forma que quem criasse os indicadores não tivesse que inserir os dados e vice-versa. Os utilizadores que possuíssem os dados, como escolas, lares, etc., poderiam aceder ao Observatório com a sua conta de técnico e inserir os dados no determinado indicador, previamente criado por um gestor. Deste modo poderá ser possível reduzir os problemas anexados á criação de indicadores.

3.4.5 - Padrões de Utilização

Por fim, antes de passar à implementação de *Web Usage Mining* no Observatório, foi necessário registar quais os parâmetros que seriam interessantes registar. Sendo uma funcionalidade nova no Observatório, as escolhas dos parâmetros foram baseadas em conversas com o orientador, que neste caso era a pessoa com mais conhecimento dos utilizadores e das pessoas envolvidas no Observatório. Os parâmetros escolhidos foram os seguintes:

- Número de utilizadores e de visualizações;
- Áreas mais visitadas;
- Indicadores mais visitados;
- Percursos dos utilizadores entre áreas e indicadores;

Traduzindo estes parâmetros para requisitos de uma ferramenta obteve-se:

- Registar os utilizadores que acedam no Observatório;
- Registar as páginas mais visualizadas
- Registar os percursos dos utilizadores

Foi então efectuada uma pesquisa por uma ferramenta que preenchesse estes requisitos. Nesta pesquisa foi possível verificar que existe um grande leque de aplicações disponíveis que permitem efectuar os mais variados registos nos servidores *Web*, inclusive os pretendidos. De forma a limitar as opções, decidiu-se que esta teria que ser livre, ser escrita em PHP e que efectuasse os seus registos em tabelas da base de dados de forma a ser mais fácil para futuras implementações e alterações. A opção recaiu sobre uma ferramenta de seu nome *phpTrafficA* [46]. O *phpTrafficA* é uma ferramenta estatística GPL para análise de tráfego *Web*, escrita em PHP e MySQL.

Depois de a ferramenta ter sido instalada e configurada foi necessário efectuar um pré-processamento dos dados para os poder apresentar no Observatório Social de Santa Maria da Feira. Essas apresentações foram alvo também de um moroso estudo, requisitando também de algum feedback por parte dos utilizadores. As apresentações de estatísticas basearam-se gráficos de barras e linhas e gráficos circulares. Para representar percursos de utilizadores recorreu-se a matrizes de adjacências e desenho de grafos.

Uma matriz de adjacência é uma das formas de representar um grafo. Tendo um grafo G com n vértices é possível representa-lo numa matriz n x n. Quanto ao desenho dos grafos, houve uma enorme dificuldade em encontrar uma ferramenta numa das linguagens usadas aqui no Observatório, que permitisse criar grafos. Como tal foi criada uma solução com base em PHP e na biblioteca PHP gráfica *gdLibrary*.

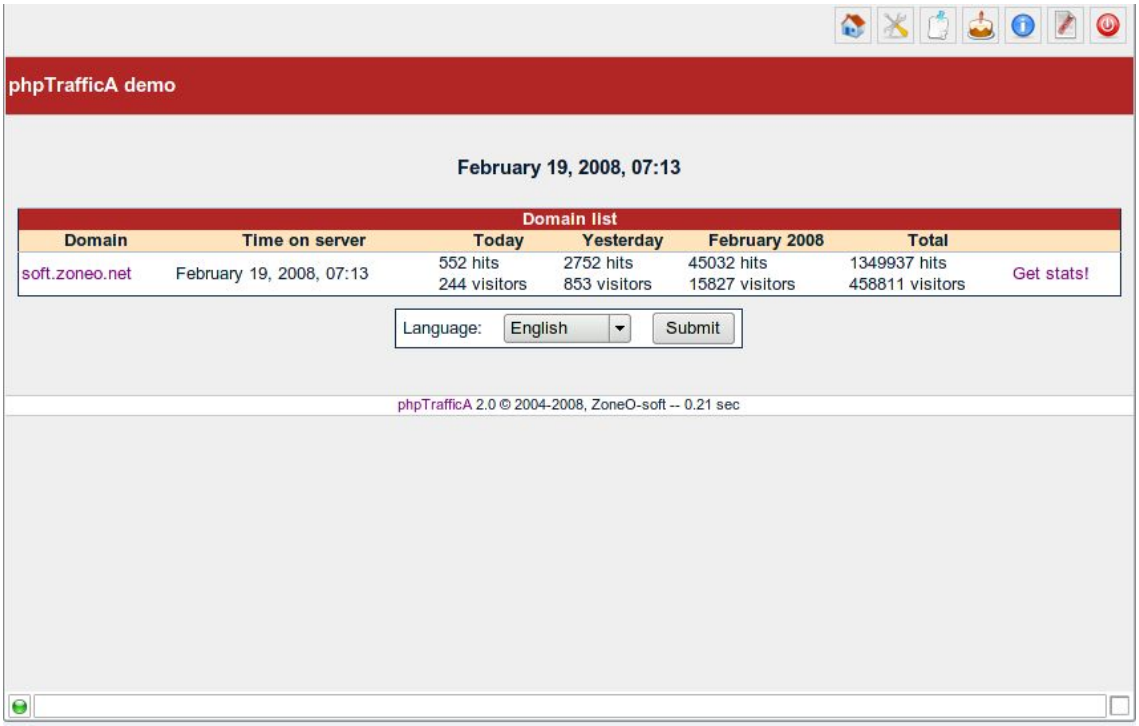


Figura 3.23: Exemplo da janela principal do *phpTrafficA* [46]

Capítulo 4

Implementação e Validação

Neste capítulo serão demonstradas as soluções implementadas para os diferentes problemas apresentados ao longo da dissertação, seguindo-se o percurso explanado no capítulo anterior. Inicialmente, são apresentadas as soluções implementadas para a criação de indicadores com quatro dimensões e para a estrutura do nome destes indicadores. Posteriormente, é apresentada a solução e os passos de implementação necessários para resolver os diferentes problemas da visualização. Segue-se a solução adoptada para a gestão de utilizadores e, para finalizar, é apresentada a implementação da análise de tráfego de utilizadores.

Como já referido no capítulo anterior, para a implementação das soluções aqui apresentadas foram usadas as seguintes linguagens: PHP, HTML, SQL, JavaScript e AJAX.

4.1 - Indicadores de Quatro Dimensões

Seguindo o percurso referido, o primeiro ponto desta implementação foi resolver o problema da estrutura dos indicadores. De forma a tornar a construção dos indicadores linear, foi feito um estudo pormenorizado do código, para que a criação dos indicadores com quatro dimensões seguisse o mesmo molde que os de menos dimensões.

O processo criação de um indicador de quatro dimensões é idêntico a um de menor dimensões, podendo ser descrito pela figura 4.1. O utilizador quando acede ao painel de administração, terá que escolher a área, o tema e o grupo que pretende para criar esse indicador. Ao entrar na página de "Criar Indicador" já possui a oportunidade de os criar com quatro dimensões, figura 4.3. Quando insere o indicador, este é adicionado à tabela "Indicadores" com os parâmetros que o constituem. Seguidamente, o utilizador poderá ir à página de "Inserir Dimensões" para escolher as dimensões que pretende para esse indicador e atribuir-lhe os diferentes valores, figura 4.2. À medida que as dimensões são criadas, vão sendo adicionadas à base de dados tabelas, com o nome da dimensão seguida do número do indicador. É também adicionada a uma outra tabela, "ind_dim", estas dimensões, a sua ordem e o indicador a que pertencem. Após a selecção das dimensões e a inserção de valores, o utilizador pode inserir os dados correspondentes aos diferentes cruzamentos de dimensões. Estes valores serão inseridos numa vista criada para conter os dados do indicador.

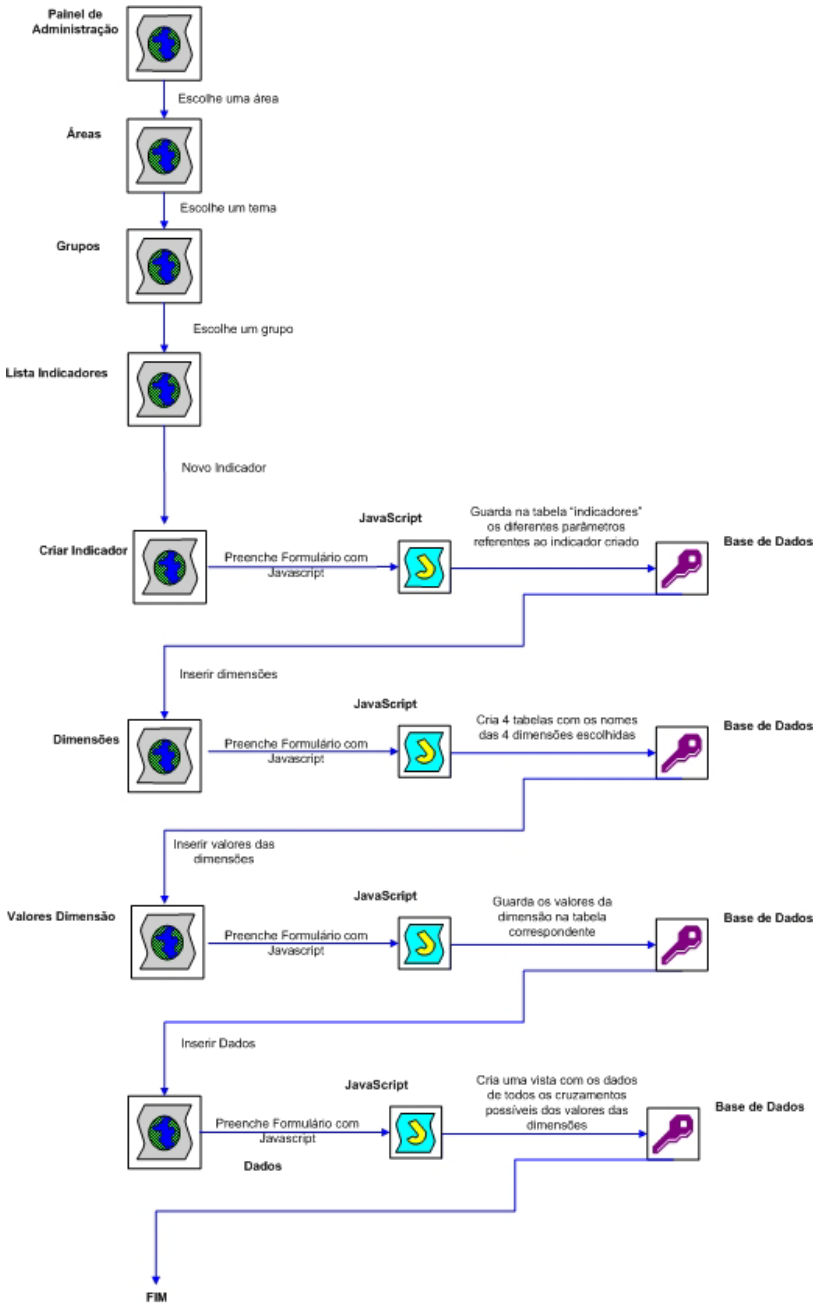


Figura 4.1: Processo simplificado de criação de um indicador de quatro dimensões

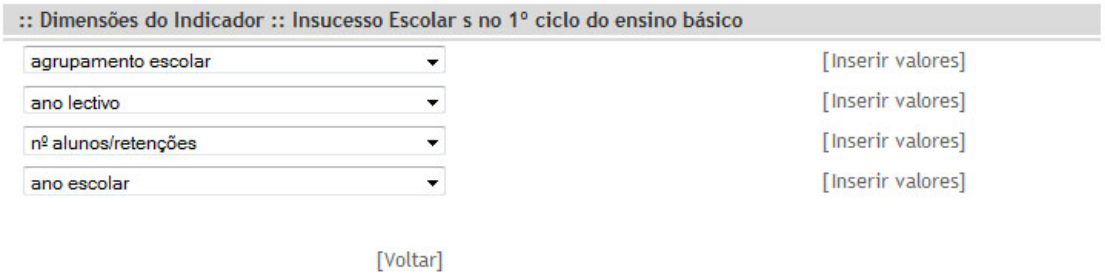


Figura 4.2: Página de inserção das dimensões do indicador

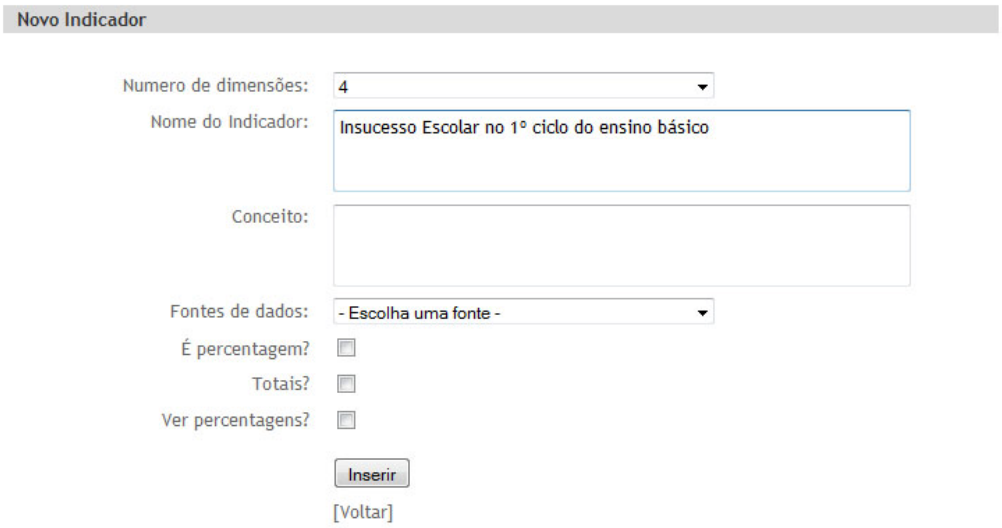


Figura 4.3: Página de criação de indicador

De modo a manter a mesma linha de pensamento operacional, independentemente da solução que viria a ser adoptada para a visualização das tabelas, foram implementadas diferentes funções para criação das tabelas de quatro indicadores, tal como eram apresentadas previamente no Observatório. O resultado é o apresentado na figura 4.4.

Insucesso Escolar s no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo, nº alunos/retenções, ano escolar.

	Ano lectivo 2004/2005								Ano lectivo 2005/2006							
	Nº alunos				Nº retenções				Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	126	167	174	164	1	20	11	26	125	144	147	181	43	10	44	3
AV Arrifana	77	88	86	75	5	28	5	8	83	113	100	112	0	14	4	6
AV Canedo	91	116	95	91	0	12	0	4	93	105	96	94	0	8	4	5
AV Fernando Pessoa	269	294	282	236					226	262	280	277	0	17	3	12
AV Fiães	119	141	142	149	0	22	5	13	156	135	124	159	0	12	8	13
AV Lobão	152	139	169	144	0	13	10	5	133	156	138	153	0	6	0	4
AV Lourosa					0	6	1	9								
AV Paços Brandão	54	57	58	61	0	32	12	14	44	60	51	58	0	3	0	6
AV Prof. Ferreira Almeida	161	195	175	206	0	12	6	2	164	160	165	174	0	20	10	4
AV Milheirós Poiares	122	124	127	126	0	16	0	11	105	135	107	125	1	9	8	14
AH Nogueira, Mozelos e Lamas	205	207	234	217					191	218	213	248	0	21	7	9

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

Figura 4.4: Exemplo de uma tabela para um indicador de quatro dimensões

Com esta implementação, foi possível reduzir de forma expressiva o número de indicadores presentes no Observatório, ficando este mais simplificado, pois tornou-se mais fácil encontrar o pretendido e as possibilidades de detectar padrões e tendências aumentou, visto agora ser possível visionar a mesma informação para as diferentes granularidades temporais, no mesmo indicador.

4.2 - Nome dos Indicadores

Tal como mencionado no capítulo anterior, se o problema está em aparecer a dimensão X no nome do indicador, a solução passa por dar a possibilidade ao utilizador de retirar a dimensão X do nome deste. Como tal, foi necessário alterar a forma como o nome do indicador era criado e dar a possibilidade ao utilizador para escolher as dimensões que queria incluir no nome do indicador.

Primeiramente, foi necessário alterar a estrutura da tabela "ind_dim", pois é nesta que a função que cria o nome dos indicadores se baseia, visto conter o nome das dimensões e a sua ordem. A esta tabela foi adicionado um campo booleano. Se este campo estiver a 1, o nome da dimensão correspondente vai aparecer no nome do indicador. Se estiver a 0, o nome não aparece. Como tal, também foi necessário alterar a função que cria o nome dos indicadores e todas as funções que interagiam com a tabela "ind_dim", desde as responsáveis pela criação de indicadores a outras funções mais simples. Por defeito, todos os valores do campo booleano estão a 1, o que significa que as dimensões correspondentes ao indicador vão aparecer no nome. Para o utilizador modificar esse valor, foi implementada na página de "Alterar Dados do Indicador" essa possibilidade, tal como está representado na figura 4.5. Por norma todas as dimensões contêm um visto, o que significa que estão presentes. Caso o utilizador desmarque uma delas, o nome da dimensão já não aparecerá, tal como está no exemplo da figura 4.6, onde o utilizador desmarcou "nº alunos/retenções".

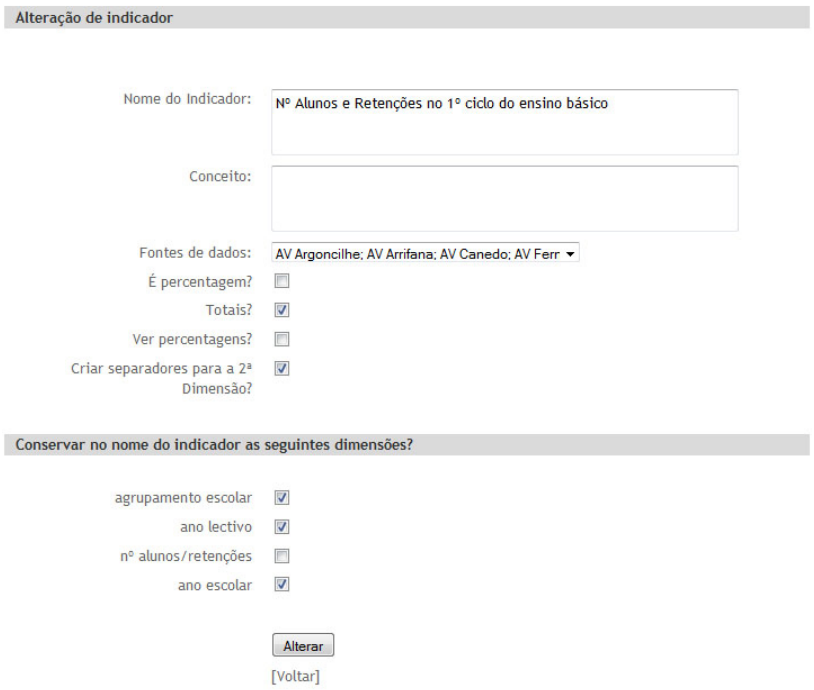
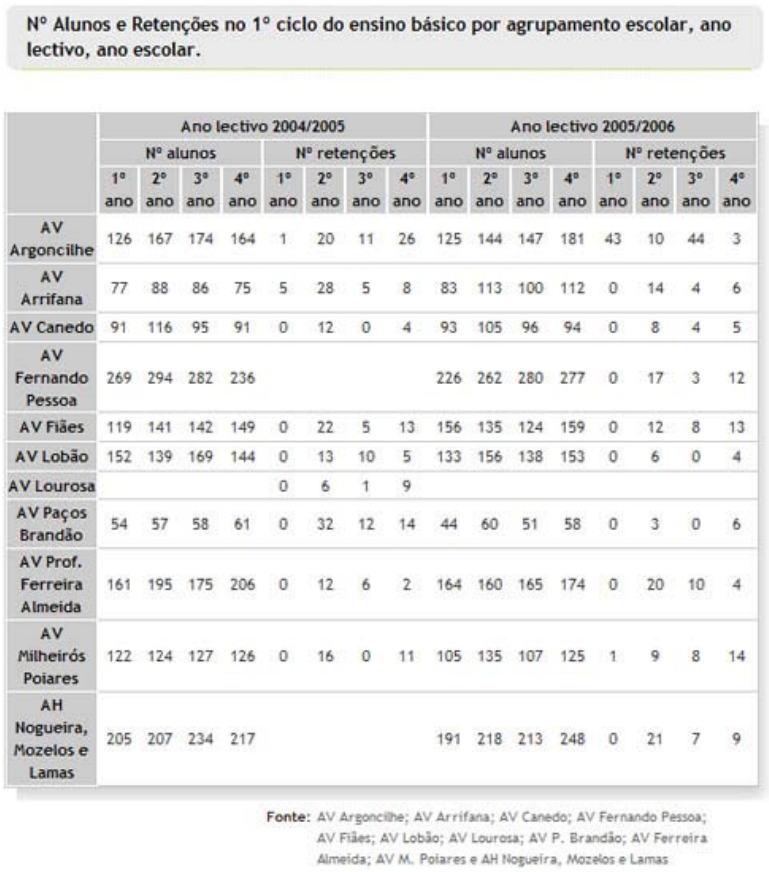


Figura 4.5: Solução adoptada para resolver o problema do nome dos indicadores



Nº Alunos e Retenções no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo, ano escolar.

	Ano lectivo 2004/2005								Ano lectivo 2005/2006							
	Nº alunos				Nº retenções				Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	126	167	174	164	1	20	11	26	125	144	147	181	43	10	44	3
AV Arrifana	77	88	86	75	5	28	5	8	83	113	100	112	0	14	4	6
AV Canedo	91	116	95	91	0	12	0	4	93	105	96	94	0	8	4	5
AV Fernando Pessoa	269	294	282	236					226	262	280	277	0	17	3	12
AV Fiães	119	141	142	149	0	22	5	13	156	135	124	159	0	12	8	13
AV Lobão	152	139	169	144	0	13	10	5	133	156	138	153	0	6	0	4
AV Lourosa					0	6	1	9								
AV Paços Brandão	54	57	58	61	0	32	12	14	44	60	51	58	0	3	0	6
AV Prof. Ferreira Almeida	161	195	175	206	0	12	6	2	164	160	165	174	0	20	10	4
AV Milheirós Poiares	122	124	127	126	0	16	0	11	105	135	107	125	1	9	8	14
AH Nogueira, Mozelos e Lamas	205	207	234	217					191	218	213	248	0	21	7	9

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

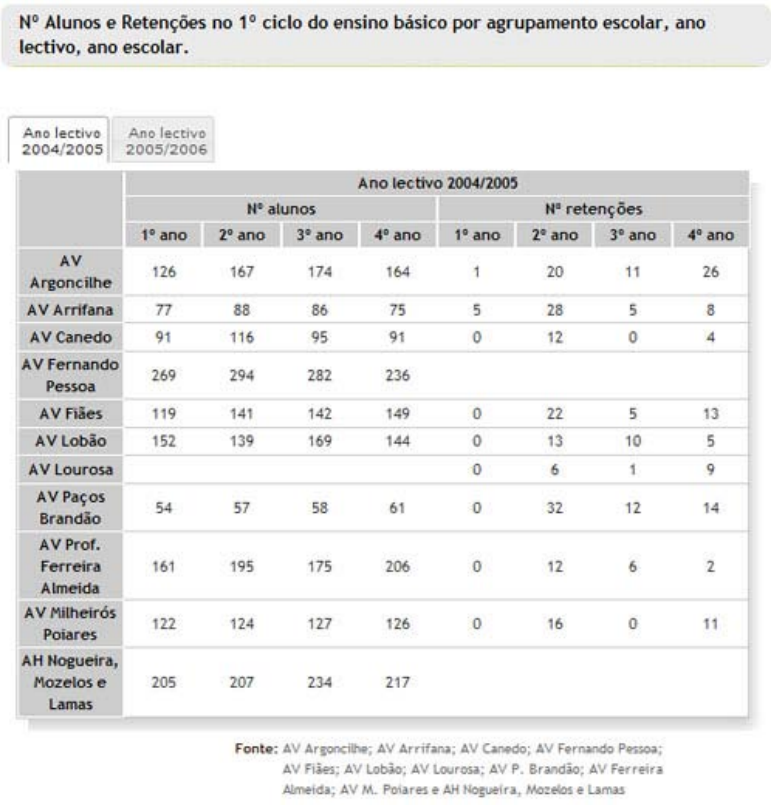
Figura 4.6: Exemplo de um indicador com a dimensão "nº alunos/retenções" omitida

4.3 - Visualização

De forma a resolver o problema da visualização em tabelas, tal como discutido anteriormente, foram implementados separadores através de JavaScript e AJAX, onde os conteúdos são apresentados dinamicamente sem a necessidade de recarregar a página. Para esta implementação recorreu-se a uma biblioteca JavaScript/AJAX, denominada *jQuery* [53].

Os separadores são criados de acordo com os valores presentes na segunda dimensão, isto é, é criado um separador para cada valor que a segunda dimensão possua, estando os dados subdivididos em várias tabelas. Foi necessário implementar várias funções distintas para suportar duas, três e quatro dimensões. No caso de indicadores só com uma dimensão, esta implementação não se justifica pois o problema da visualização não ocorre, já que os valores da primeira dimensão são apresentados verticalmente.

Um exemplo da solução está representada na figura 4.7. Esta tabela é a aplicação de separadores para a figura 4.6. Como o separador do ano lectivo 2004/2005 está seleccionado (aparece a branco), a tabela que aparece é aquela que possui os valores para esse ano lectivo. Se o utilizador pretender visualizar um outro ano lectivo apenas tem que clicar no respectivo separador, figura 4.8.



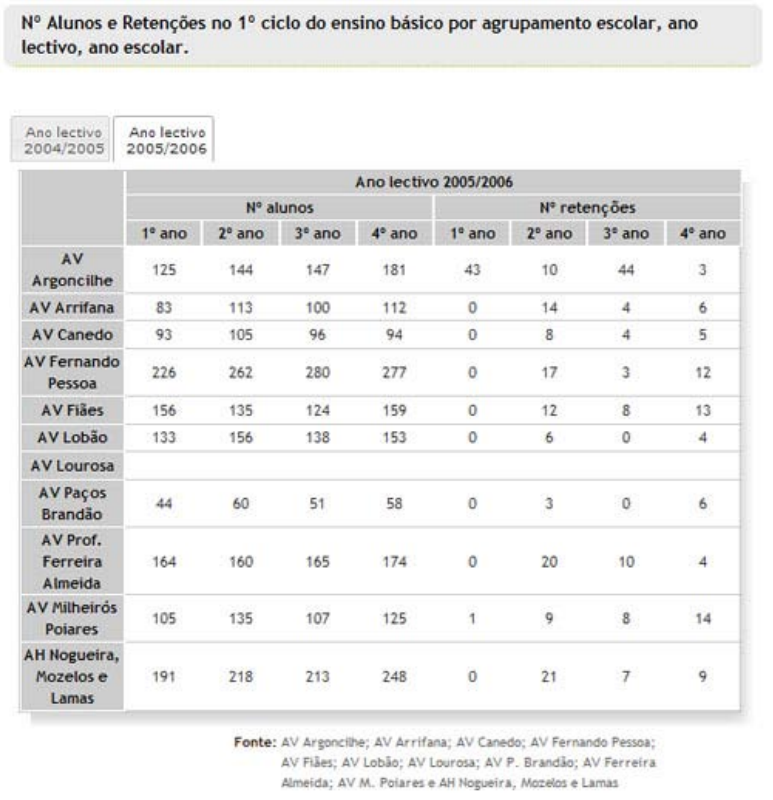
Nº Alunos e Retenções no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo, ano escolar.

Ano lectivo 2004/2005 | Ano lectivo 2005/2006

	Ano lectivo 2004/2005							
	Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	126	167	174	164	1	20	11	26
AV Arrifana	77	88	86	75	5	28	5	8
AV Canedo	91	116	95	91	0	12	0	4
AV Fernando Pessoa	269	294	282	236				
AV Fiães	119	141	142	149	0	22	5	13
AV Lobão	152	139	169	144	0	13	10	5
AV Lourosa					0	6	1	9
AV Paços Brandão	54	57	58	61	0	32	12	14
AV Prof. Ferreira Almeida	161	195	175	206	0	12	6	2
AV Milheirós Poiares	122	124	127	126	0	16	0	11
AH Nogueira, Mozelos e Lamas	205	207	234	217				

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

Figura 4.7: Visualização de um indicador para o ano lectivo 2004/2005



Nº Alunos e Retenções no 1º ciclo do ensino básico por agrupamento escolar, ano lectivo, ano escolar.

Ano lectivo 2004/2005 | Ano lectivo 2005/2006

	Ano lectivo 2005/2006							
	Nº alunos				Nº retenções			
	1º ano	2º ano	3º ano	4º ano	1º ano	2º ano	3º ano	4º ano
AV Argoncilhe	125	144	147	181	43	10	44	3
AV Arrifana	83	113	100	112	0	14	4	6
AV Canedo	93	105	96	94	0	8	4	5
AV Fernando Pessoa	226	262	280	277	0	17	3	12
AV Fiães	156	135	124	159	0	12	8	13
AV Lobão	133	156	138	153	0	6	0	4
AV Lourosa								
AV Paços Brandão	44	60	51	58	0	3	0	6
AV Prof. Ferreira Almeida	164	160	165	174	0	20	10	4
AV Milheirós Poiares	105	135	107	125	1	9	8	14
AH Nogueira, Mozelos e Lamas	191	218	213	248	0	21	7	9

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

Figura 4.8: Visualização de um indicador para o ano lectivo 2005/2006

Contudo, a implementação dos separadores poderá não se justificar em todos os casos. Por exemplo, quando a dimensão temporal não está presente ou quando a segunda dimensão possui apenas um valor, o uso de separadores torna-se irrelevante. Como tal é também necessário que a visualização de tabelas esteja preparada para a opção que dispense o uso de separadores.

Assim, para solucionar esta questão, foi dada a opção de escolha ao utilizador para a criação de separadores, tal como está representado na figura 4.9. Caso pretenda criar separadores deve assinalar a opção "Criar separadores para a segunda Dimensão" com um visto. De referir que na implementação das funções responsáveis por criar as tabelas com separadores, foram também adicionados os parâmetros opcionais (percentagens e totais), de forma a enriquecer o conteúdo e a informação para o utilizador.

Foi apenas feita uma alteração que se achou relevante. De facto, constatou-se que, no Observatório, quando o utilizador assinalava com um visto a opção "Totais?", na visualização das tabelas surgia a soma horizontal e vertical dos dados, que nem sempre não fazia sentido. Este caso encontra-se evidenciado na figura 4.10, onde as colunas verticais "Total" representam as somas do nº de aluno com o nº de retenções para as respectivas escolas. Assim, optou-se por retirar essas colunas e apenas deixar a soma das colunas verticais.

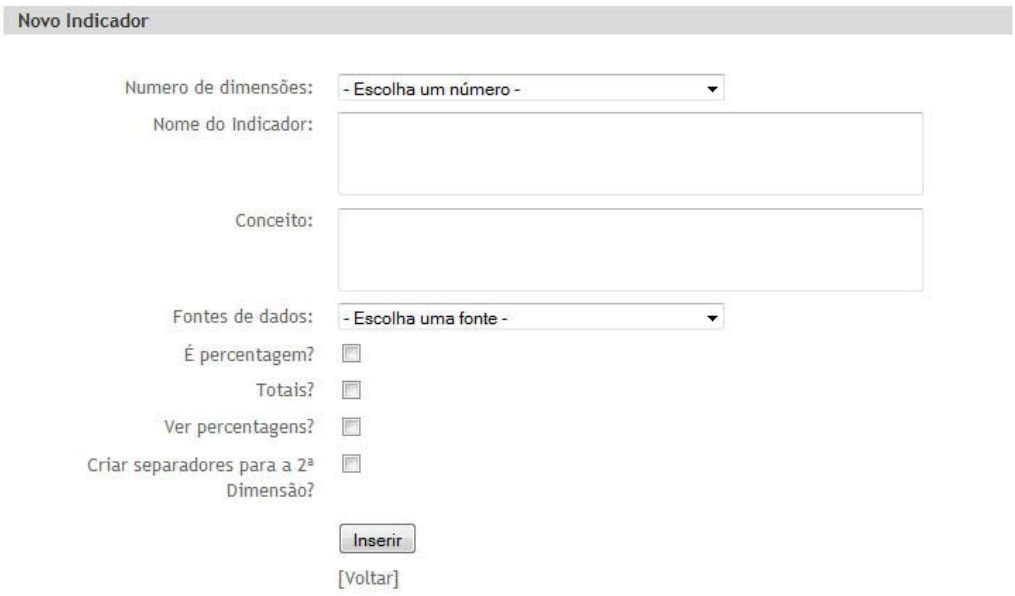


Figura 4.9: Opção "Criar separadores para a segunda Dimensão" em Novo indicador

Insucesso escolar no 3º ciclo do ensino básico por agrupamento escolar/esc. sec. Stª Mª Feira/ colégio liceal Stª Mª Lamas, ano lectivo e nº alunos/retenções

	Ano lectivo 2004/2005					Ano lectivo 2005/2006					
	Nº alunos		Nº retenções		Total	Nº alunos		Nº retenções		Total	Total
AV Argoncilhe	339	6.9%	12	2.1%	351 6.4%	355	7.2%	64	10%	419 7.5%	770 7%
AV Arrifana	291	5.9%	77	13.8%	368 6.7%	278	5.6%	40	6.2%	318 5.7%	686 6.2%
AV Canedo	324	6.6%	36	6.4%	360 6.6%	312	6.3%	50	7.8%	362 6.5%	722 6.5%
AV Fernando Pessoa	507	10.4%	-	-	507 9.3%	569	11.5%	82	12.8%	651 11.7%	1158 10.5%
AV Fiães	447	9.1%	93	16.6%	540 9.9%	423	8.6%	84	13.1%	507 9.1%	1047 9.5%
AV Lobão	394	8%	52	9.3%	446 8.2%	414	8.4%	58	9%	472 8.5%	918 8.3%
AV Lourosa	-	-	50	8.9%	50 0.9%	-	-	-	-	0 0%	50 0.5%
AV P. Brandão	476	9.7%	59	10.6%	535 9.8%	437	8.8%	66	10.3%	503 9%	1038 9.4%
AV Prof. Ferreira Almeida	428	8.7%	-	-	428 7.8%	505	10.2%	30	4.7%	535 9.6%	963 8.7%
AV M. Poiares	368	7.5%	98	17.5%	466 8.5%	351	7.1%	80	12.5%	431 7.7%	897 8.1%
Escola Secundária Stª Mª Feira	365	7.5%	-	-	365 6.7%	342	6.9%	19	3%	361 6.5%	726 6.6%
Colégio Liceal Stª Mª Lamas	959	19.6%	82	14.7%	1041 19.1%	957	19.4%	68	10.6%	1025 18.4%	2066 18.7%
Total	4898	100%	559	100%	5457 100%	4943	100%	641	100%	5584 100%	11041 100%

Figura 4.10: Exemplo de uma tabela com os totais

Após a implementação da solução para o problema da visualização em tabelas, o trabalho centrou-se na visualização gráfica.

A solução para este problema, baseia-se na interacção entre o utilizador e as tabelas. O primeiro passo foi implementar as funções que constroem os gráficos, de acordo com as escolhas dos utilizadores. Uma vez que os indicadores contêm diferentes dimensões, foi necessário trabalhá-las separadamente. A visualização de indicadores de uma dimensão não foi necessário modificar, pois a que estava presente no Observatório era a adequada.

Começando pelos gráficos de dois indicadores, a solução adoptada encontra-se representada na figura 4.11. Este gráfico apresenta o número de alunos nas diferentes escolas, no ano lectivo 2004/2005. Aqui a primeira dimensão é o "agrupamento escolar" e a segunda é o "ano lectivo". Para construir esta visualização, o utilizador teria que ter seleccionado a segunda dimensão, mais propriamente, o ano lectivo 2004/2005.

Para os indicadores de três dimensões, os gráficos já vão conter uma modificação. Se no caso anterior adicionarmos uma terceira dimensão para possuir o número de alunos e o número de reprovações então o gráfico vai ser o da figura 4.12. Neste caso, seguindo a ordem que os valores da dimensão apresentam na tabela, as colunas a azul representam o número de alunos e as colunas a verde representam o número de reprovações. Se a terceira dimensão possuísse mais valores, o gráfico iria ter mais colunas de cores diferentes em cada agrupamento escolar.

Quanto aos indicadores de quatro dimensões, de forma a conseguir-se representar uma quarta dimensão, utilizaram-se gráficos de colunas empilhadas. Seguindo ainda o exemplo anterior, dividindo o nº de alunos e o nº de retenções pelos respectivos anos lectivos, o gráfico gerado é o representado pela figura 4.13.

A selecção de cores a ser usada foi alvo de grande discussão e de variados testes e estudos. A este nível, é muito importante conseguir que os utilizadores, ao analisar os gráficos, consigam distinguir os valores e detectar algum tipo de padrão e/ou tendência.

De salientar, que a representação gráfica de indicadores com quatro dimensões possui uma legenda, de maneira a que os utilizadores percebam rapidamente o que representa e a não surgir dúvidas.

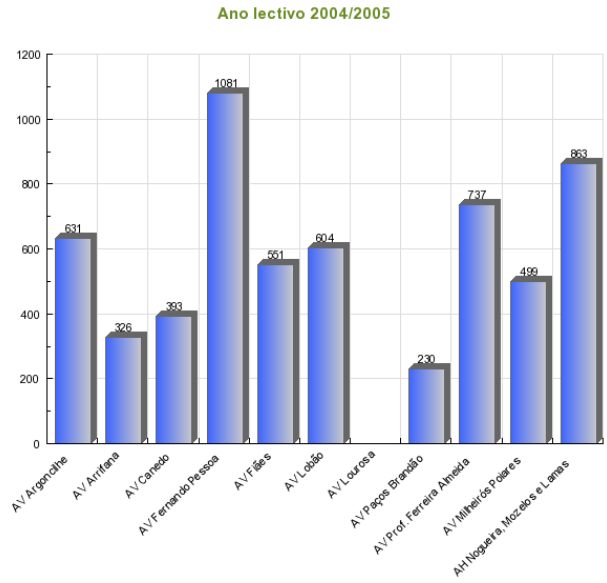


Figura 4.11: Exemplo de gráfico para um indicador de duas dimensões

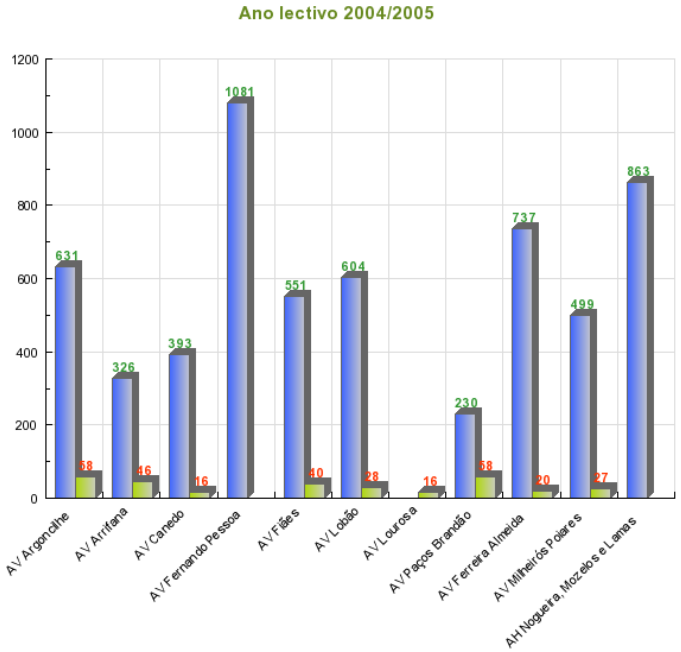


Figura 4.12: Exemplo de gráfico para um indicador de três dimensões

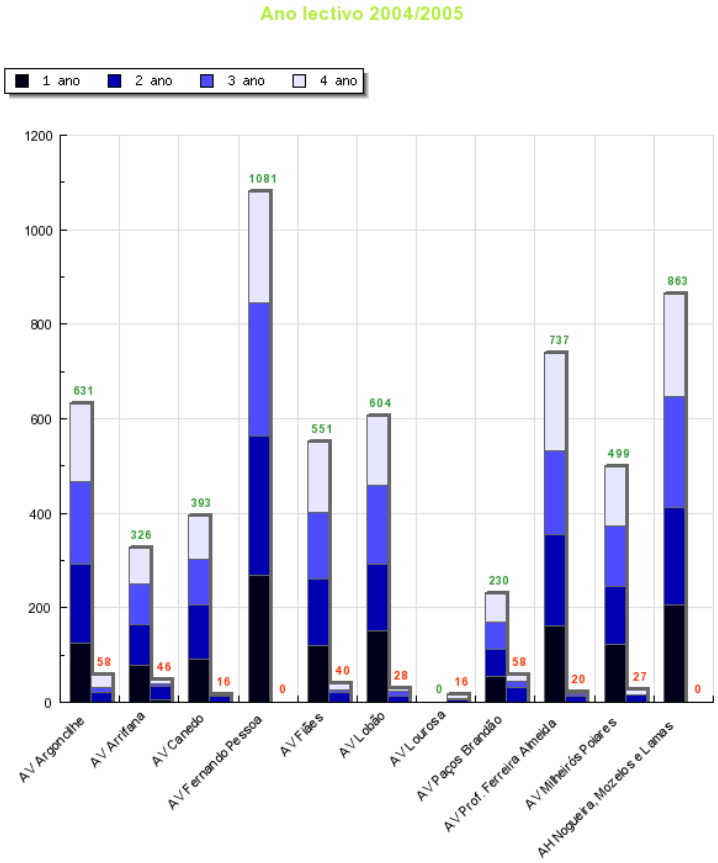


Figura 4.13: Exemplo de gráfico para um indicador de quatro dimensões

Uma vez implementadas as soluções gráficas previamente descritas, inclusive através de parâmetros recebidos, foi necessário tornar a tabela interactiva. Esses parâmetros são submetidos pelos utilizadores quando clicam num valor da primeira ou segunda dimensão. Esta interacção foi conseguida através de JavaScript e AJAX. Com JavaScript criaram-se os botões para os valores da primeira e segunda dimensão. Através de AJAX foi possível enviar a escolha do utilizador para o *script* responsável por gerar e guardar o gráfico para, de seguida, apresentá-lo; isto tudo ocorre de um modo rápido e sem recarregar a página.

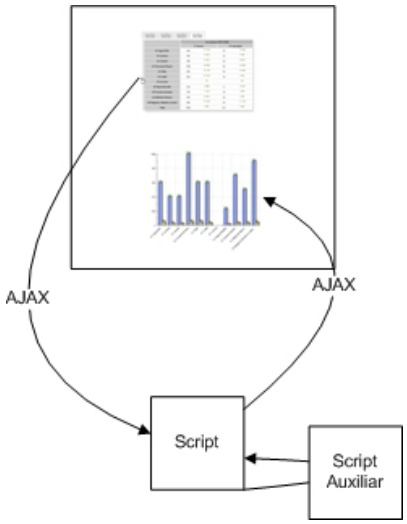


Figura 4.14: Ligação tabela - gráfico

Um exemplo da solução pode ser observado nas figuras 4.15 e 4.16. Em 4.15 o utilizador clicou no agrupamento escolar "AV Fiães", sendo lhe apresentado o gráfico com os dados desse agrupamento, para os diferentes valores da segunda dimensão, neste caso, para os diferentes anos lectivos. Por sua vez, em 4.16, o utilizador clicou no "Ano Lectivo 2007/2008", sendo-lhe apresentados os dados de todos os agrupamentos escolares para esse ano.

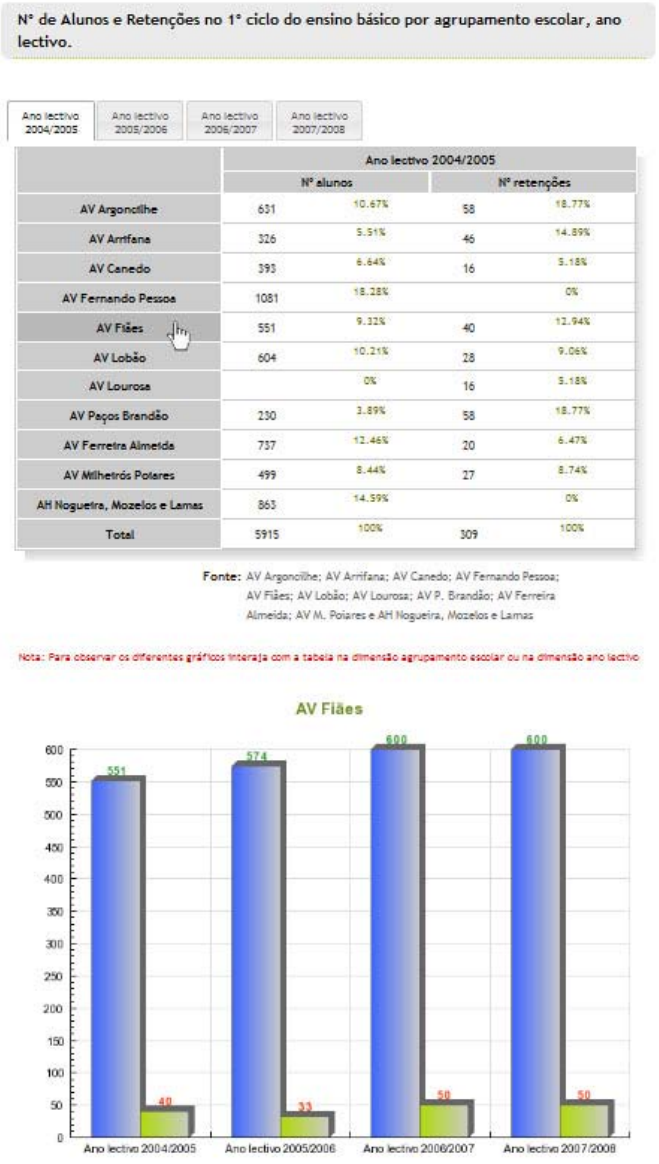


Figura 4.15: Solução Final - Visualização primeira dimensão

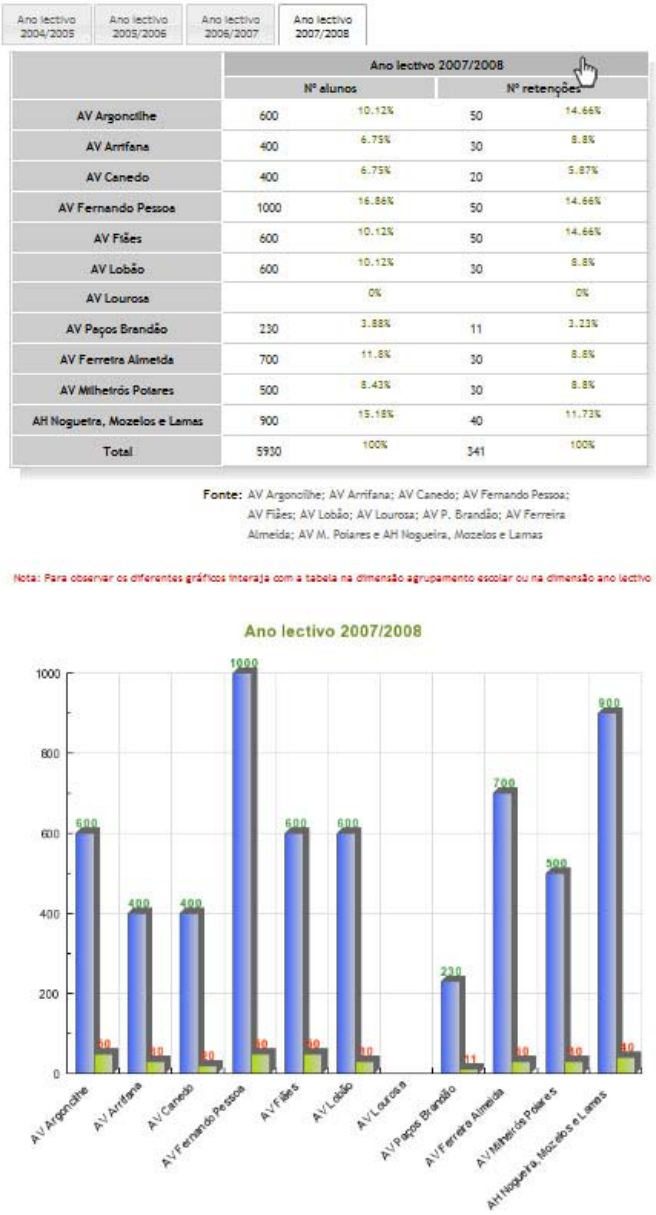


Figura 4.16: Solução Final - Visualização segunda dimensão

Havia agora outro tipo de visualização gráfica que necessitava de ser revista. A visualização cartográfica do concelho de Santa Maria da Feira. O Observatório já possuindo este tipo de visualização, apenas necessitava que fosse acrescentada a possibilidade de no mesmo indicador apresentar os mapas para os diferentes registos temporais, ou seja, criar uma terceira dimensão. Assim, em vez de serem criados vários indicadores com essa informação, seria criado apenas um. Este tipo de representação visual, só era possível com indicadores de duas dimensões, facto colmatado com o presente trabalho, que permite a implementação de três dimensões.

Adaptando também aqui as tabelas dinâmicas, desenvolveu-se a possibilidade do utilizador clicar nos diferentes valores da segunda dimensão, sendo-lhe apresentados os respectivos mapas. Para esta implementação foi necessário ignorar as funções já criadas no Observatório para este efeito e criar novas funções com diferentes estruturas.

Em conversas com os próprios utilizadores do Observatório, foi possível constatar que neste tipo de visualização, os utilizadores sentiam carência de uma legenda. Os utilizadores apenas conseguiam analisar os valores por freguesia se analisassem as tabelas, pois o gráfico não possuía qualquer legenda para a paleta cromática utilizada. Foi então acrescentada às funções responsáveis pela criação dos mapas, códigos para gerar estas legendas de acordo com os dados. Desta forma, os utilizadores não necessitam de possuir as tabelas para ter uma ideia geral dos valores representados pelos mapas.

Ainda na visualização em mapas, o modo como os intervalos de cor eram gerados, não era o ideal. O que as funções faziam era achar o valor máximo e o mínimo, subtraia-os e dividia esse resultado por cinco. Eram criados então cinco intervalos desde o valor mínimo até ao máximo. A freguesia com o valor máximo iria conter a cor mais "escura" não existindo mais nenhuma com essa mesma cor, nem que tivesse apenas menos uma unidade. Com algumas pesquisas efectuadas, foi possível afirmar que tal não era o mais correcto pois iria alterar como os utilizadores vêm esses dados e isolar freguesias tornando-as como referência. Outra dúvida ocorrida é que estes intervalos podiam conter valores decimais.

O usual neste tipo de visualizações é promover intervalos simples de serem observados. Na grande maioria dos exemplos que foi possível analisar durante a investigação preliminar, estes intervalos são sempre constituídos por valores inteiros e de fácil leitura, numa proporcional, coerente e dependente dos valores máximos. Deste modo, a forma como os intervalos no Observatório eram gerados, foi modificada. Resumidamente, o que o algoritmo faz é o seguinte:

- Percorre os dados à procura do valor máximo;
- O valor máximo é dividido pelo número de intervalos pretendidos, que neste caso são cinco;
- Se esse resultado for inteiro significa que os intervalos serão formados por múltiplos de cinco logo numa óptica racional, são valores aceitáveis;
- Se o resultado não for inteiro os valores não são considerados aceitáveis, logo o valor máximo para gerar o mapa e a legenda é somado de uma unidade e feita a divisão por cinco;
- Caso o resultado seja agora inteiro os valores já são considerados aceitáveis e gerar-se-á o mapa e a legenda a partir desse valor.

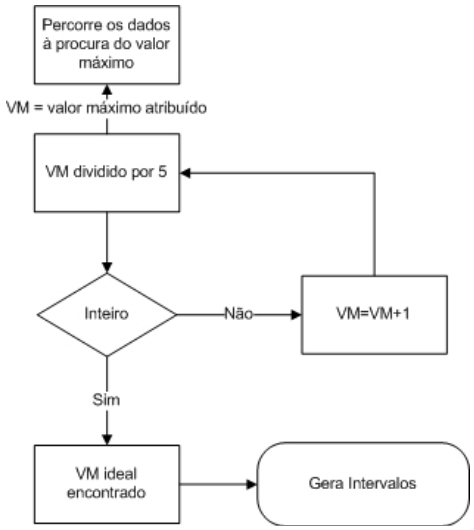
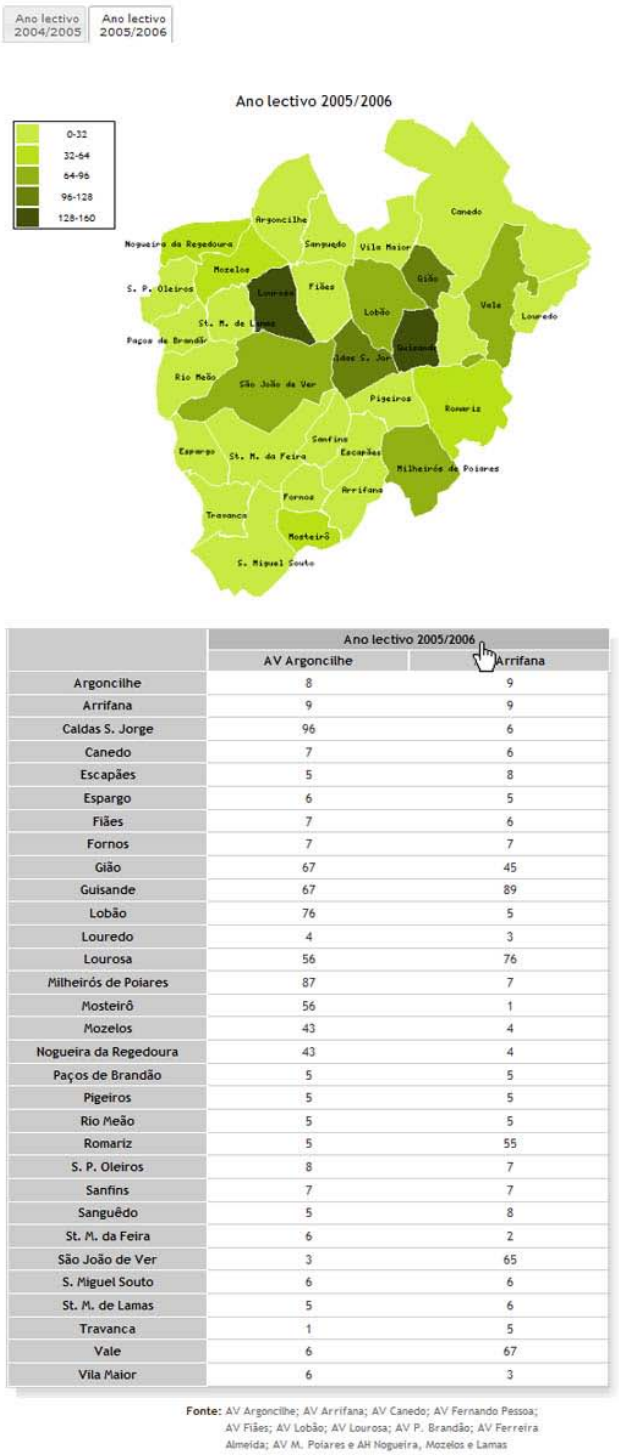


Figura 4.17: Diagrama de blocos do algoritmo que gera os intervalos

O número de intervalos escolhidos foi de cinco, pois na literatura de referência, o valor considerado óptimo é cinco ou seis, devido à capacidade de percepção dos utilizadores. É usado o método dos intervalos iguais para os definir, isto é, para calcular o tamanho do intervalo, divide-se o valor máximo por cinco. De facto, existem vários métodos para determinar os intervalos, mas como se trata de um mapa pequeno, com apenas trinta e uma áreas, resolveu-se usar o mais simples e perceptível para a maioria dos utilizadores, [35][36].



	Ano lectivo 2005/2006	
	AV Argoncilhe	Arrifana
Argoncilhe	8	9
Arrifana	9	9
Caldas S. Jorge	96	6
Canedo	7	6
Escapães	5	8
Espargo	6	5
Fiães	7	6
Fornos	7	7
Gião	67	45
Guisande	67	89
Lobão	76	5
Louredo	4	3
Lourosa	56	76
Milheirós de Poiares	87	7
Mosteirô	56	1
Mozelos	43	4
Nogueira da Regedoura	43	4
Paços de Brandão	5	5
Pigeiros	5	5
Rio Meão	5	5
Romariz	5	55
S. P. Oleiros	8	7
Sanfins	7	7
Sanguêdo	5	8
St. M. da Feira	6	2
São João de Ver	3	65
S. Miguel Souto	6	6
St. M. de Lamas	5	6
Travanca	1	5
Vale	6	67
Vila Maior	6	3

Fonte: AV Argoncilhe; AV Arrifana; AV Canedo; AV Fernando Pessoa; AV Fiães; AV Lobão; AV Lourosa; AV P. Brandão; AV Ferreira Almeida; AV M. Poiares e AH Nogueira, Mozelos e Lamas

Figura 4.18: Exemplo da solução num indicador com a visualização em mapas

4.4 - Gestão de Utilizadores

Como foi referido no capítulo anterior, as permissões no painel administrativo foram distribuídas por três tipos de utilizadores: administrador, gestor e técnico.

O primeiro passo foi alterar a estrutura de login do painel, para quando um utilizador efectuar o login, seja reconhecido o seu tipo e lhe sejam dadas as devidas permissões. Tendo agora a necessidade de criar e apagar novos utilizadores, foi criada uma página de gestão de utilizadores à qual apenas o administrador tem acesso. Nesta página, o administrador pode criar novos utilizadores, ou então procurar por utilizadores existentes e apagá-los, figura 4.19.

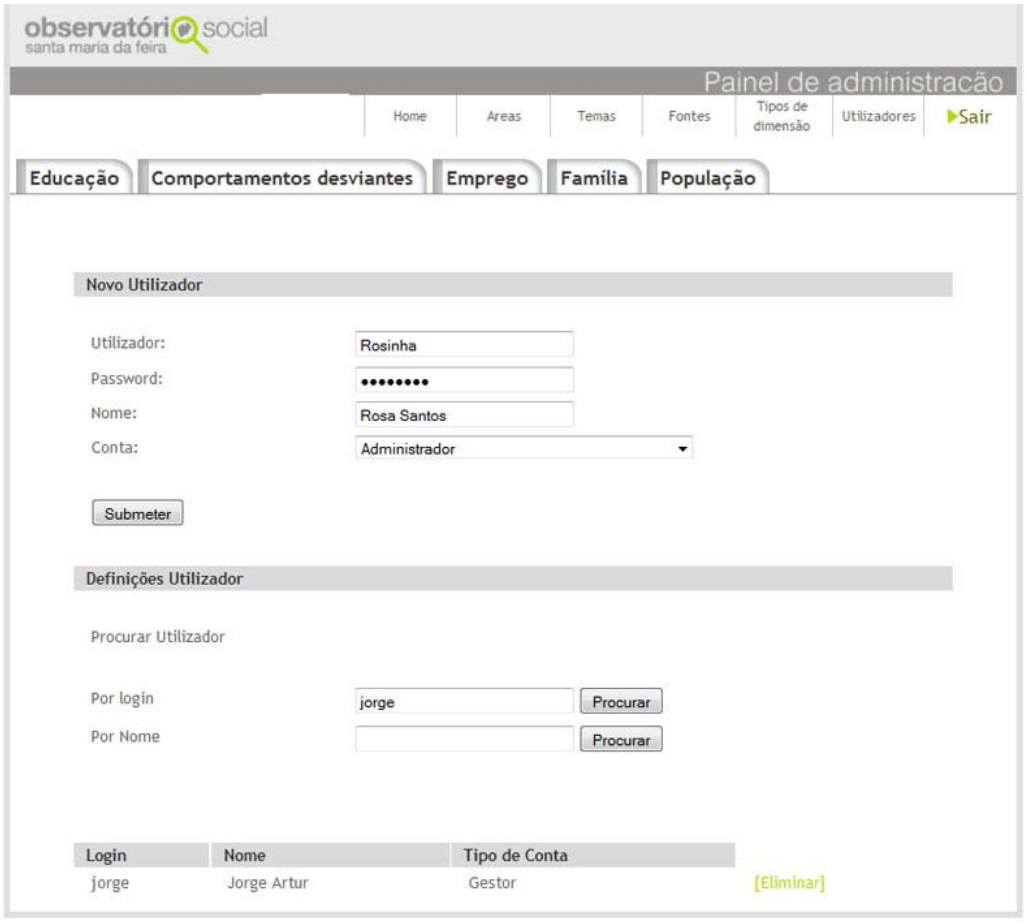


Figura 4.19: Página de administração de utilizadores

Com a inserção de vários utilizadores, achou-se relevante criar uma página para os utilizadores poderem configurar as suas contas. Por agora apenas possui a possibilidade de alterar a palavra-chave, figura 4.20. Implementado isto, apenas faltava implementar as restrições a cada utilizador.

Figura 4.20: Página de administração de conta

As permissões implementadas para cada utilizador são as seguintes:

Administrador

- Criar/Alterar/Apagar Áreas, Temas, Fontes, Tipos de Dimensão, Utilizadores, Indicadores;
- Inserir dados nos indicadores;
- Configuração de conta;

Gestor

- Criar/Alterar/Apagar Áreas, Temas, Fontes, Tipos de Dimensão, Indicadores;
- Inserir dados nos indicadores;
- Configuração de conta;

Técnico

- Inserir dados nos indicadores;
- Configuração de conta;

4.5 - Padrões de Utilização

De forma a registar os padrões de utilização, tal como foi descrito no capítulo anterior, foi instalada uma ferramenta de análise de tráfego denominada *phpTrafficA* [46]. A sua instalação consistiu, numa primeira parte, em configurar alguns dos ficheiros PHP; como por exemplo, configurar o acesso à base de dados de forma a guardar os registos. De seguida extraíram-se os ficheiros da ferramenta para o servidor do Observatório e correu-se o *script Install.php* num *software* de navegação, seguindo as instruções que iam aparecendo.

Tendo a plataforma instalada, acedeu-se à página principal da ferramenta no servidor. Aqui, o primeiro passo foi configurar o domínio pretendido para a realização dos registos, figura 4.21. Nesta dissertação, como não era possível trabalhar directamente no servidor real, o domínio usado foi o próprio IP do computador que estava a ser usado. Contudo, como este

se ia alterando, foi usado também o próprio domínio (*Localhost)*, para efectuar os diferentes testes e configurações ao longo desta implementação. De forma a iniciar-se os registos no domínio pretendido, apenas foi necessário preencher alguns campos, tal como a figura 4.21 ilustra. Os campos obrigatórios são o endereço de domínio *(Domain Address)* e nome base das tabelas *(Table name)* que são criadas com os diferentes registos. Existem depois outros campos que são extremamente importantes para o adequado funcionamento desta ferramenta, num servidor público.

O campo *Public* indica se todos os utilizadores têm acesso à visualização dos registos deste domínio, ou se apenas tem acesso quem fizer o respectivo *login*. Como não se pretende que todos os utilizadores tenham acesso a estes registos, pelas mais variadas razões, como privacidade, informações confidenciais, entre outras, a opção ficou como privada. Outro campo é o *Trim URL*. Se este estiver como verdadeiro os endereços vão ser cortados, por exemplo, os endereços http://obsocial.idit.up.pt/index.php?m=3 e http://obsocial.idit.up.pt/index.php?m=4 vão ser ambos registados como http://obsocial.idit.up.pt/index.php. Como isto não é o desejado, pois o Observatório é um *Website* dinâmico com base em http://obsocial.idit.up.pt/index.php, sendo necessário fazer registos individuais de páginas, este campo ficou como falso.

Existem mais dois campos. O primeiro questiona se se pretende registar ou ignorar os programas de rastreio, como o *googlebot*, e o segundo permite a introdução de um contador de visitas, nas diferentes páginas *Web*. No âmbito deste trabalho, ambos os campos ficaram como *falsos*.

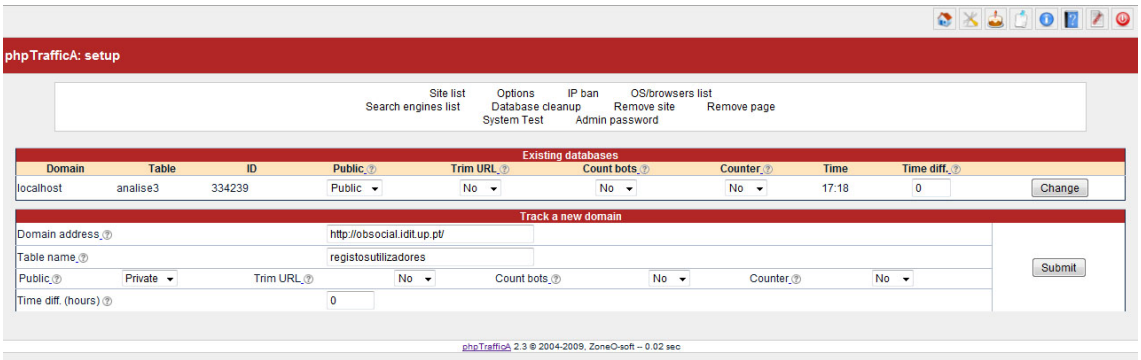


Figura 4.21: Página de configurações do *phpTrafficA*

Tendo o domínio configurado, apenas faltava um passo para iniciar os registos. Colocar um pequeno código na página principal do Observatório, neste caso, o *index.php.* Este código foi obtido na página "*Reminder*" da ferramenta - quarto ícone do canto superior direito da figura 4.21.

O problema de não estar num servidor acessível a todos os utilizadores limitou o estudo e os testes a serem efectuados. Os registos demonstrados ao longo desta implementação foram com base em acessos e percursos efectuados aleatoriamente, sem qualquer tipo de tendência, por um ou mais utilizadores, em dias diferentes. Contudo, como o ambicionado era a implementação da ferramenta e de formas de visualização para demonstrar alguns dos dados de utilização aos comuns utilizadores, estes registos eram suficientes.

O *phpTrafficA* efectua os seus registos em tabelas da base de dados. Para cada domínio, este cria quinze tabelas diferentes com vários tipos de registos: acessos, *programas* de navegação usados, países, utilizadores, IPs, páginas, percursos, tempos de retenção, entre

outros. Com base nestes registos, a ferramenta oferece ao administrador um variado leque de visualizações como visitas diárias, semanais, mensais, páginas mais vistas desde o inicio dos registos, ou páginas mais vistas do mês X ou semana Y, ou do dia Z. As figuras 4.22, 4.23 e 4.24 evidenciam alguns exemplos de visualização que podem ser encontrados nesta ferramenta.

Figura 4.22: Exemplo de visualização no *phpTrafficA*: acessos e páginas mais populares da semana

Figura 4.23: Exemplo de visualização no *phpTrafficA*: Top 10

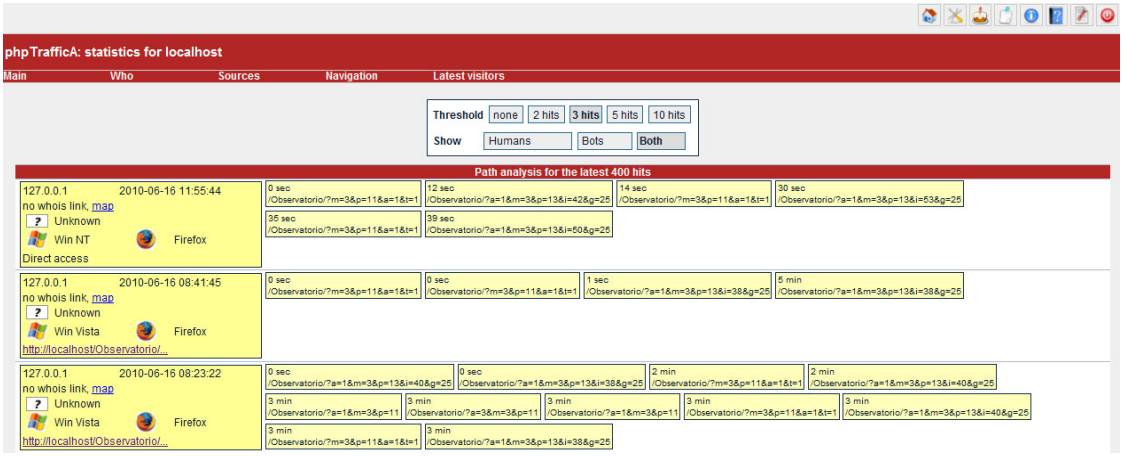


Figura 4.24: Exemplo de visualização no phpTrafficA: Percurso dos Utilizadores

Embora existam várias opções de visualização nesta ferramenta, estas não permitem aceder a uma análise concreta e facilmente entendível da navegação feita pelo utilizador no Observatório. Através destas visualizações não é possível obter os requisitos pré-definidos para esta dissertação, como saber as áreas e indicadores mais visualizados. Para além disso, estas visualizações só estão acessíveis aos administradores e nunca aos utilizadores comuns. Um dos objectivos que presidiu à realização do presente trabalho, era disponibilizar algumas informações aos utilizadores. De forma a cumpri-lo, foram analisadas todas as tabelas com os registos efectuados por esta ferramenta.

A implementação das diferentes visualizações foi dividida em duas partes. Inicialmente, foi trabalhada uma página com as visualizações sobre acessos, áreas e indicadores. Posteriormente, foi trabalhada outra página com os percursos dos utilizadores.

Em ambas as páginas, os registos apresentados são mensais, tendo sido implementada a possibilidade dos utilizadores realizarem uma análise por mês de interesse. Para o efeito, começou-se por implementar na primeira página um formulário para o utilizador escolher a data na qual pretende observar os dados, tendo todas as futuras funções sido criadas para receber esses parâmetros.

4.5.1 - Visualizações de Utilização

De seguida, iniciaram-se as implementações das visualizações. A primeira a ser efectuada foi o número de utilizadores e o número de acessos a páginas do Observatório. Esta implementação conseguiu-se através de duas tabelas de registos. Uma fornece o número de utilizadores por dia no *Website* e nas diferentes páginas. A outra fornece o número total de visualizações no servidor e em diferentes páginas deste. Extraída esta informação, foi criada a visualização gráfica para apresenta-la. Neste caso a visualização escolhida foi um gráfico de barras e de linhas, figura 4.25.

Passando à segunda visualização gráfica, o pretendido era apresentar as áreas mais visualizadas. Para tal, foi necessário efectuar uma formatação dos dados dos registos. Para obter os acessos às diferentes páginas e os seus respectivos endereços, foi necessário interligar duas tabelas através de SQL, pois a tabela que fornece os acessos às diferentes páginas fornece apenas a sua referência e não o endereço destas. Possuindo os endereços das páginas e o seu número de visualizações, foi implementado um algoritmo para formatação dos endereços. Sendo os endereços do tipo http://obsocial.idit.up.pt/index.php?, o único modo

de saber se este corresponde a uma determinada área é através do parâmetro "a" no endereço, isto é, observando a tabela 4.1, o 1º e 3º endereços possuem o parâmetro "a", 3 e 1 respectivamente, o que significa que pertencem à terceira e primeira área. Relativamente ao 2º endereço, como não possui o parâmetro, não corresponde a nenhuma área. De uma forma simplificada, o algoritmo deste processo é o representado pela figura 4.26.

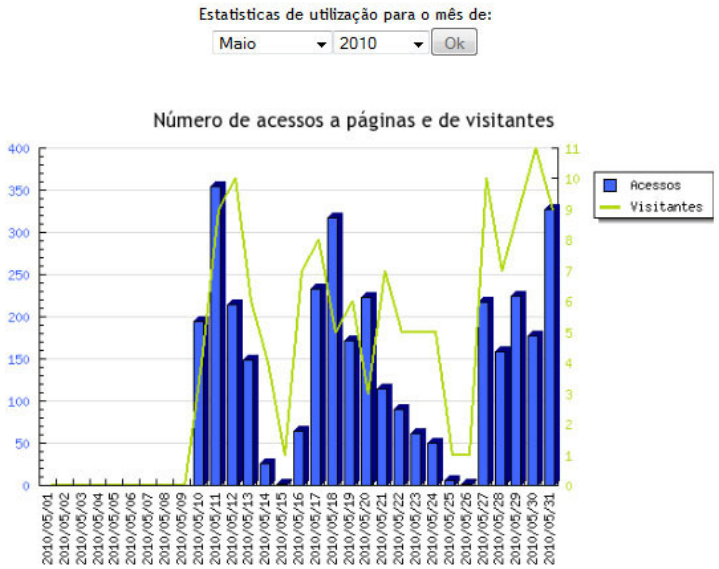


Figura 4.25: Número de acessos a páginas e de visitantes

Tabela 4.1: Exemplo de Endereço e Área correspondente

Endereço	Área
http://obsocial.idit.up.pt/index.php?a=3&m=4	3
http://obsocial.idit.up.pt/index.php?p=2	NULL
http://obsocial.idit.up.pt/index.php?g=1&a=1	1

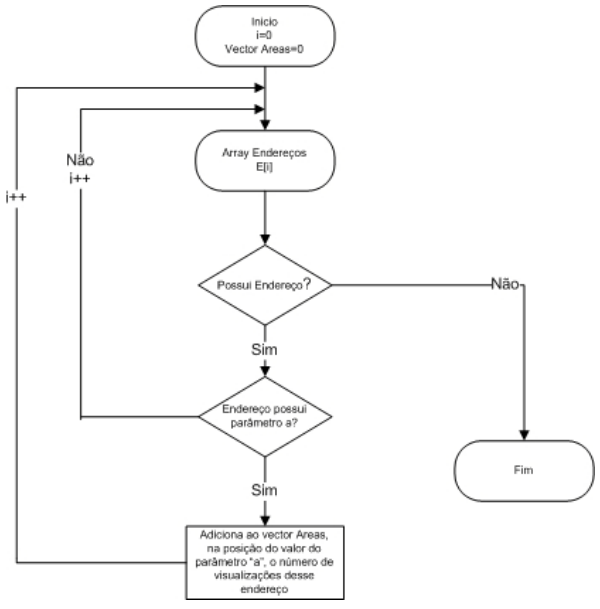


Figura 4.26: Fluxograma do algoritmo para cálculo da visualização das áreas

Uma vez possuindo o vector relativo ao número de visualizações de cada área, antes de poder construir a visualização, e de forma a construir a sua legenda, foi necessário extrair os nomes das áreas da tabela da base de dados que possui essa informação. Um exemplo da visualização implementada é apresentado na figura 4.27.

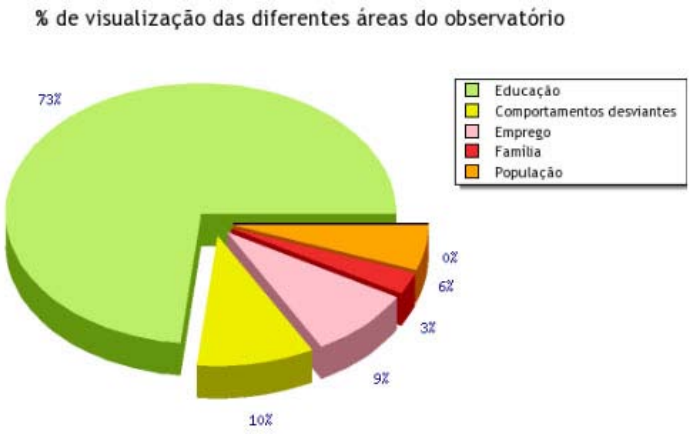


Figura 4.27: % de visualização das diferentes áreas do Observatório

Após se ter implementado a visualização das áreas, transitou-se para a dos indicadores. Aqui, o pretendido era demonstrar a percentagem dos dez indicadores mais visualizados e o número de visualizações dos cinco indicadores menos visualizados. Para este objectivo foi também necessária uma formatação dos dados. De forma idêntica ao processo efectuado para as áreas, aqui o parâmetro não seria o "a" mas sim o "i". Este indica qual o indicador que corresponde a esse endereço. Exemplificando, na tabela 4.2 o 1º e 2º endereços possuem o parâmetro "i", logo correspondem a um indicador. Já o 3º exemplo não possui esse parâmetro, logo não corresponde a um indicador, não contando para a visualização.

Tabela 4.2: Exemplo de Endereço e Indicador correspondente

Endereço	Indicador
http://obsocial.idit.up.pt/index.php?m=3&p=13&i=593&g=111	593
http://obsocial.idit.up.pt/index.php?m=3&p=13&i=383&g=84	383
http://obsocial.idit.up.pt/index.php?m=3&p=11&a=1&t=1	NULL

O algoritmo responsável por esta formatação é muito idêntico ao implementado para as áreas, como é possível notar pela figura 4.28. O *Array* Endereços vai possuir os vários endereços e o número respectivo de visualizações. Esta extracção vai também depender da visualização a efectuar. No caso dos indicadores mais visualizados, o *Array* Endereços vai estar ordenado de forma descendente. Se a visualização pretendida for a dos menos visualizados a ordem vai ser ascendente.

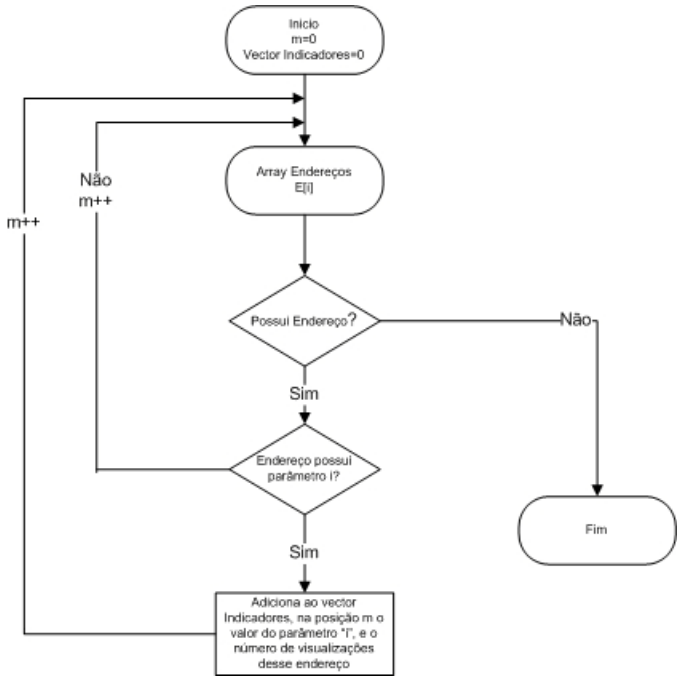


Figura 4.28: Fluxograma do algoritmo para o cálculo da visualização dos indicadores

Depois de mais algumas formatações e restrições, foi possível gerar os gráficos tendo como exemplos as figuras 4.29 e 4.30.

Figura 4.29: Os indicadores mais visualizados e respectivas %

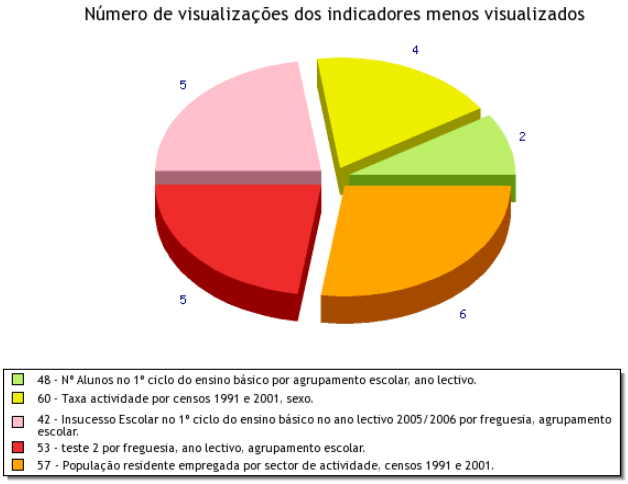


Figura 4.30: Número de visualizações dos indicadores menos visualizados

4.5.2 - Percursos

Após a conclusão da implementação das diferentes visualizações de utilização, deu-se início à implementação da visualização dos percursos. O objectivo é apresentar os percursos efectuados pelos utilizadores, de indicador para indicador. Como tal, era necessário estudar as tabelas que registavam os percursos, figura 4.31. A análise desta tabela permitiu constatar que iria ser necessária uma enorme formatação dos dados. Esta vai registando os percursos efectuados, assinalando as transições de página para página, ou para a mesma página, através das referências criadas, pela ferramenta, para cada página. Com isto, o primeiro passo foi transformar todos os percursos presentes na tabela num *Array* de vectores.

Por exemplo, o percurso |40|21|23|45|3| transformar-se-ia em Percurso[0]=(40,21,23,45,3), o percurso |1|3|2|43| transformar-se-ia em Percurso[1]=(1,3,2,43), seguindo este princípio até não existir mais percursos na tabela.

Como os percursos guardados na tabela são séries de caracteres, foi necessário implementar um algoritmo capaz de extrair as referências separadas por os caracteres "|".

Após possuir o *Array* de vectores com todos os percursos, implementou-se outro algoritmo para retirar as páginas que não fossem indicadores. O algoritmo verifica a que endereço corresponde a referência e extrai-o. Uma vez possuindo o endereço, verifica se este é indicador e a que área este corresponde, aplicando os mesmos algoritmos referidos anteriormente. Caso não seja indicador, essa página é ignorada e avança para a seguinte referência. Caso seja indicador, é guardada a sua referência e a da área em que ele está inserido, passando de seguida à próxima referência.

Após possuir as transições todas no novo *Array*, este é sujeito a um novo algoritmo, que vai guardando numa tabela da base de dados os saltos efectuados de indicador para indicador. Caso existam saltos dentro do mesmo indicador, esses serão ignorados e não contarão para a estatística. Este foi um processo bastante demorado devido à complexidade da estrutura e à forma como o *phpTrafficA* guardava os registos. Foi necessário efectuar várias *queries* e várias formatações.

Um diagrama muito simplificado com os passos deste processo pode ser observado na figura 4.33. Um exemplo da tabela gerada com as diferentes transições de indicadores ocorridos pode ser consultada na figura 4.32. Esta tabela contém o número de ocorrências das transições entre indicadores, as áreas a que eles pertencem e a data do registo desses saltos.

A data foi inserida para o utilizador poder observar esses saltos mensalmente, como será explicado de seguida.

	id	first	last	entry	exit	path	length	count
	1	2010-06-01 08:22:17	2010-06-01 08:22:17	49	82	|49|49|49|49|50|49|49|49|50|49|49|50|50|75|50|50|5...	70	1
	2	2010-06-01 09:55:07	2010-06-01 09:55:07	49	50	|49|49|49|50|83|49|50|50|83|50|82|50|50|74|50|77|5...	145	1
	3	2010-06-01 11:49:35	2010-06-01 11:49:35	23	35	|23|2|53|2|35|35|35|35|	8	1
	4	2010-06-01 11:49:35	2010-06-01 11:49:35	23	35	|23|2|53|2|35|35|35|35|	8	1
	5	2010-06-01 17:31:11	2010-06-01 17:31:11	86	3	|86|86|35|22|23|2|53|2|1|2|2|3|3|3|3|2|4|4|4|4|2...	106	1
	6	2010-06-01 19:34:17	2010-06-01 19:34:17	3	5	|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3|3...	63	1
	7	2010-06-02 08:23:41	2010-06-02 08:23:41	2	2	|2|4|24|25|25|2|22|6|49|49|23|25|25|2|	14	1
	8	2010-06-02 15:45:06	2010-06-02 15:45:06	27	101	|27|6|49|49|23|2|4|2|4|6|49|22|23|2|4|6|6|49|50|10...	28	1
	9	2010-06-02 20:08:25	2010-06-02 20:08:25	49	27	|49|27|27|	3	1
	10	2010-06-02 20:47:11	2010-06-02 20:47:11	23	103	|23|23|2|103|	4	1
	11	2010-06-02 21:28:55	2010-06-02 21:28:55	103	103	|103|	1	1
	12	2010-06-10 16:41:49	2010-06-10 16:41:49	27	103	|27|52|52|42|51|23|2|103|2|103|2|103|103|	13	1
	13	2010-06-10 17:17:19	2010-06-10 17:17:19	2	104	|2|2|104|	3	1
	14	2010-06-14 09:25:56	2010-06-14 09:25:56	27	3	|27|23|23|2|4|2|3|3|	8	1
	15	2010-06-14 11:22:37	2010-06-14 11:22:37	3	3	|3|3|3|3|3|	5	1
	16	2010-06-14 18:07:50	2010-06-14 18:07:50	2	3	|2|2|3|3|	4	1
	17	2010-06-14 19:24:29	2010-06-15 08:10:31	3	3	|3|3|	2	3
	18	2010-06-15 08:47:50	2010-06-15 20:32:29	3	3	|3|	1	3
	19	2010-06-15 18:00:09	2010-06-15 18:00:09	2	1	|2|1|	2	1
	20	2010-06-15 18:25:54	2010-06-15 18:25:54	2	4	|2|2|4|2|4|2|3|2|4|4|2|3|11|12|11|11|28|28|29|28|2...	40	1
	21	2010-06-15 21:38:32	2010-06-15 21:38:32	4	3	|4|3|	2	1
	22	2010-06-15 21:38:32	2010-06-15 21:38:32	4	3	|4|3|	2	1
	23	2010-06-16 08:23:22	2010-06-16 08:23:22	3	4	|3|2|3|23|9|23|2|3|2|4|	10	1
	24	2010-06-16 08:41:45	2010-06-16 08:41:45	2	4	|2|2|4|4|	4	1
	25	2010-06-16 09:14:46	2010-06-16 09:35:09	4	4	|4|4|	2	2
	26	2010-06-16 11:55:44	2010-06-16 11:55:44	2	35	|2|5|2|26|2|35|	6	1
	27	2010-06-16 21:00:07	2010-06-16 21:00:07	35	49	|35|6|49|	3	1
	28	2010-06-17 08:36:42	2010-06-17 08:58:53	49	49	|49|49|	2	2
	29	2010-06-17 09:30:29	2010-06-17 09:30:29	23	23	|23|	1	1

Figura 4.31: Exemplo de alguns registos de percursos na tabela “nome_path”

	id	dearea	paraarea	deind	paraind	data	count
	666	3	2	59	61	2010-04-10	1
	665	3	3	60	56	2010-04-10	1
	664	2	3	64	60	2010-04-10	1
	663	2	2	61	64	2010-04-10	2
	662	3	3	57	55	2010-04-10	1
	661	2	3	61	57	2010-04-10	1
	660	2	1	61	38	2010-04-10	1
	659	1	2	50	61	2010-04-10	1
	658	1	1	53	50	2010-04-10	2
	657	2	1	61	50	2010-04-10	1
	656	1	3	40	55	2010-04-10	1
	655	1	1	53	38	2010-04-10	1
	654	1	1	50	53	2010-04-10	3
	653	1	1	48	50	2010-04-10	1
	652	3	3	59	60	2010-04-10	1
	651	3	3	57	59	2010-04-10	1
	650	1	3	50	57	2010-04-10	2
	649	3	1	59	50	2010-04-10	1
	648	3	3	60	59	2010-04-10	3
	647	3	3	57	60	2010-04-10	2
	646	3	3	56	57	2010-04-10	2
	645	3	3	55	56	2010-04-10	1
	644	3	3	56	55	2010-04-10	1
	643	1	3	40	56	2010-04-10	1
	642	1	1	42	40	2010-04-10	2
	641	1	1	39	42	2010-04-10	1
	640	1	1	40	50	2010-04-10	1
	639	1	1	38	39	2010-04-10	2
	638	1	1	38	50	2010-04-10	2
	637	1	1	50	38	2010-04-10	2

Figura 4.32: Exemplo da tabela que regista os saltos entre indicadores

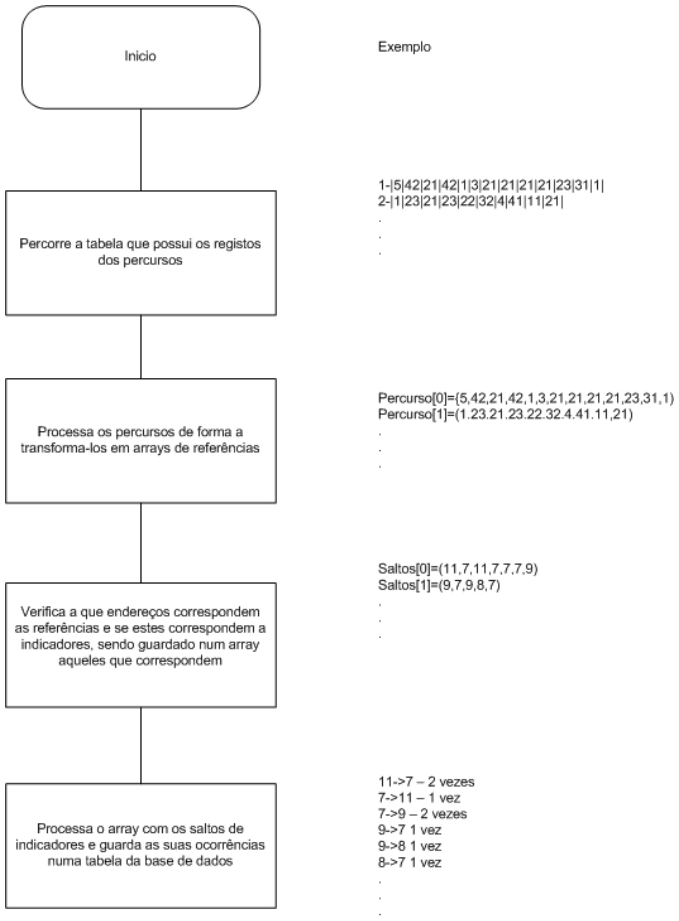


Figura 4.33: Exemplo simplificado dos passos do processo de formatação dos dados de percursos

Este processo sendo muito demorado, não faria sentido estar constantemente a ser gerado, pois sempre que um utilizador entra irá ser criado um novo percurso. Como o objectivo é observar mensalmente os trajectos, faz sentido apenas correr o processo uma vez no início de cada mês. O problema é que os registos efectuados pela ferramenta *phpTrafficA* em relação aos percursos não possuem a data de ocorrência de cada percurso. Se determinado percurso acontece várias vezes, apenas se consegue saber a primeira e ultima vez que este aconteceu. Com isto, foi implementado um *script* que corre apenas uma vez no inicio de todos os meses para copiar os registos efectuados do mês anterior pela ferramenta numa nova tabela, limpando de seguida a tabela da ferramenta de forma a iniciar novos registos. Desta forma, o processo da figura 4.33 correrá a partir desta nova tabela, gerando uma outra com as transições e as datas correspondentes, figura 4.32.

Tendo a tabela com as transições e as suas datas de ocorrência, o próximo passo consistiu na implementação das formas de visualização. Estas foram divididas em duas partes. Numa primeira parte são apresentadas as transições entre indicadores de áreas diferentes e numa segunda parte as transições entre indicadores da mesma área.

As formas de visualização aqui representadas foram implementadas através de PHP e da sua biblioteca - *gdLibrary*. Para as transições entre indicadores de áreas diferentes a visualização criada foi a representada na figura 4.34.

Nesta visualização, cada círculo representa uma área do Observatório e as rectas as transições entre áreas. O tamanho dos círculos vai depender do número de transições

ocorridas entre indicadores, da área que esse círculo representa. A espessura das linhas vai depender do número de ocorrências entre indicadores de áreas diferentes.

Observando o exemplo da figura 4.34, a área que possui mais ocorrências entre indicadores é a da Educação, pois é a que apresenta um círculo maior. Por outro lado, as áreas Família e População possuem escassas ou mesmo nenhuma ocorrência. Quanto às linhas, as transições mais ocorridas de uma área para a outra ocorreram entre Emprego e Educação. Para as áreas Família e Educação, como não existem linhas, significa que ninguém transitou para estas ou de estas, após ter estado numa outra área.

Os tamanhos dos círculos e a espessura das linhas seguem uma escala logarítmica, pois foi a maneira mais adequada encontrada para representar as variações de tamanho.

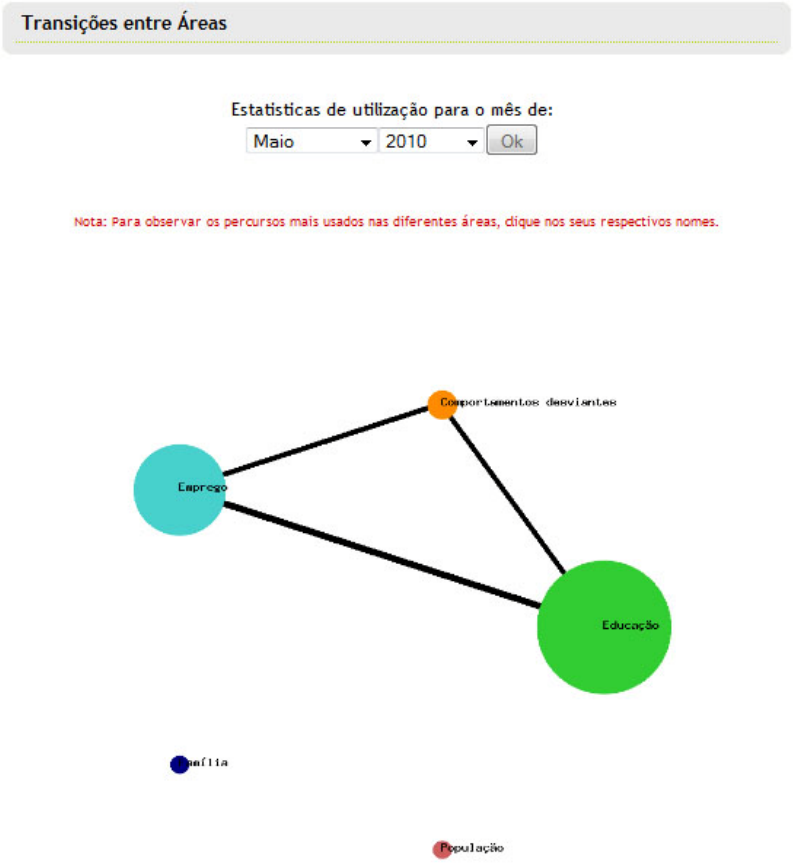


Figura 4.34: Exemplo de visualização das transições entre áreas

Para visualizar as transições entre indicadores da mesma área, tornou-se a visualização das áreas interactiva. O utilizador ao clicar no círculo da área que deseja observar, será encaminhado para a página de observação de transições dessa área. Para esta visualização, foram implementadas duas estratégias.

A primeira baseia-se no método da matriz de adjacência, figura 4.36. Esta matriz foi tornada interactiva, para que quando o utilizador passe com o cursor nos números dos indicadores, seja possível observar que indicador corresponde a este número. Optou-se por não se colocar directamente os nomes nas tabelas devido aos seus tamanhos e ao espaço reduzido que as tabelas possuem.

A outra estratégia implementada foi a visualização em grafos, figura 4.37. Devido à escassez de ferramentas em PHP ou JavaScript que permitissem a construção de grafos, esta forma de visualização no Observatório foi criada de raiz apenas com *PHP* e com a sua biblioteca - *gdLibrary*. O algoritmo implementado baseia-se na seguinte ideia: este desenha a primeira transição ocorrida com os grafos; de seguida, analisa a origem da segunda transição. Se a origem desta for igual ao destino da transição anterior, continua a desenhar o percurso no mesmo grafo. Caso seja diferente, inicia um novo percurso desenhando essa transição. O algoritmo repete-se até ter desenhado a ultima transição.

Para um exemplo mais prático inseriu-se a matriz da tabela 4.3, produzindo o resultado da figura 4.35. O algoritmo desenha a primeira transição, ou seja, 60 para 59. De seguida, analisa a origem da segunda transição, que neste caso corresponde ao índice 1. Como a origem deste não é igual ao destino da transição anterior, desenha uma nova transição abaixo desta. Analisa a próxima transição, a de índice 2 e como a origem desta é igual ao destino da transição anterior, desenha esta transição no mesmo percurso. Tal como está representado na figura 4.35. O algoritmo repete-se até ter desenhado a ultima transição.

Tabela 4.3: Matriz inserida no exemplo prático do algoritmo dos grafos

Índice	Origem	Destino	Ocorrências
0	60	59	3
1	56	57	2
2	57	60	2
3	56	55	1
4	55	56	1
5	57	59	1
6	59	60	1
7	57	55	1

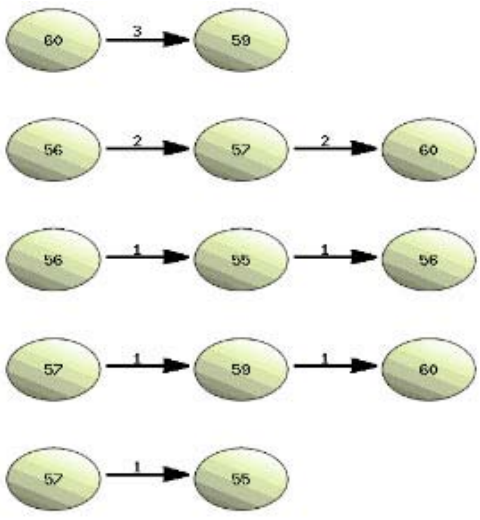


Figura 4.35: Grafo gerado pelo algoritmo dos grafos quando inserida a matriz da tabela 4.3

Os grafos foram também tornados interactivos, para que quando o utilizador desloque o cursor por cima dos nós, lhe seja apresentado os nomes correspondentes. Esta implementação poderá não ser a mais adequada, pois existiam várias relações que poderiam ser feitas entre

os grafos. Face à limitação temporal que caracteriza este tipo de trabalho de investigação, esta apresentação gráfica não foi implementada de forma a explorar todas as suas potencialidades, ficando em aberto futuros desenvolvimentos.

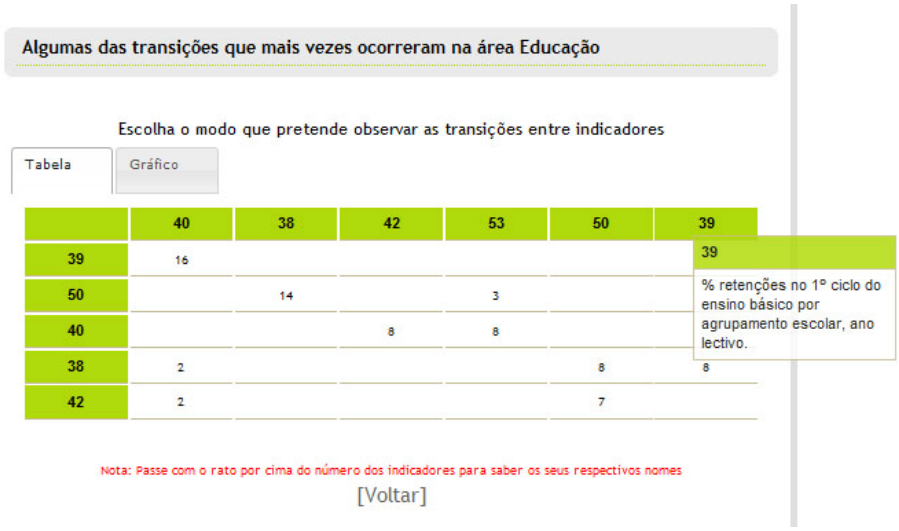


Figura 4.36: Matriz de adjacência para a visualização das transições

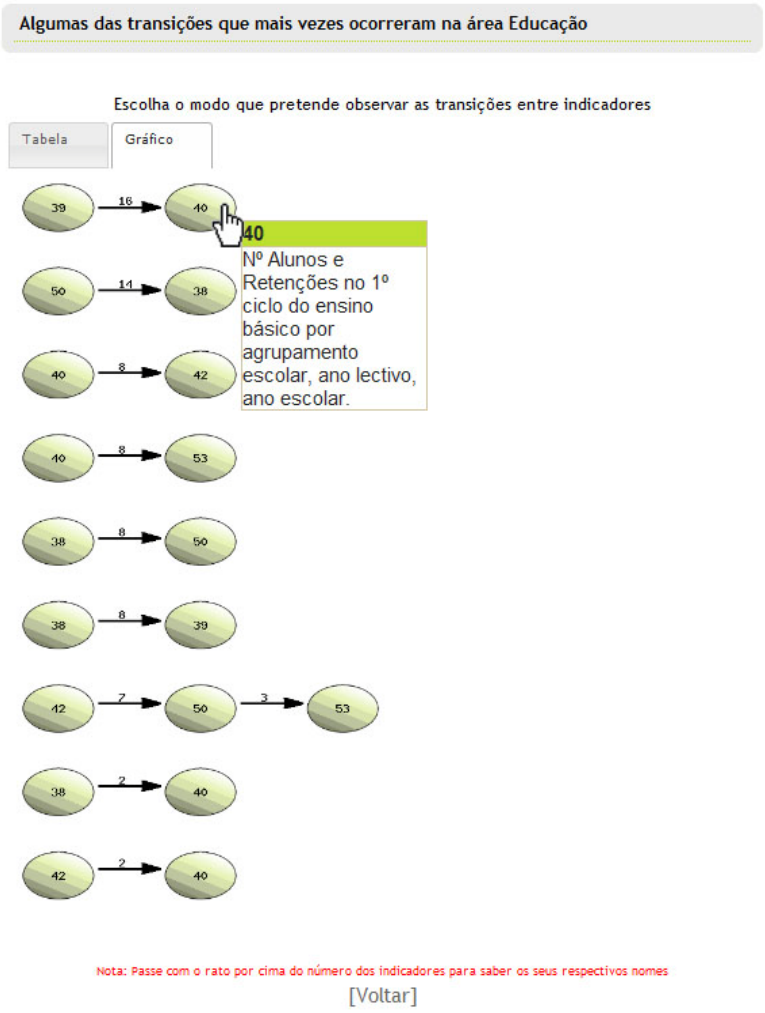


Figura 4.37: Visualização em grafos para representação das transições

Capítulo 5

Conclusões e Futuros Desenvolvimentos

5.1 - Conclusões

Tendo por base os objectivos formulados no Capítulo 1, e que surgiram para dar resposta a uma necessidade real do Observatório Social de Santa Maria da Feira, os estudos realizados ao longo desta dissertação permitiram retirar algumas conclusões importantes sobre todo o processo que engloba a construção de um observatório deste género. Permitiram também retirar variadas conclusões sobre a visualização de grandes quantidades de informação, dados espaços temporais e de métodos para extracção e análise de padrões de utilização na *Web*.

A construção de um observatório como o trabalhado nesta dissertação pode ser um processo extremamente complicado. Este necessita de guardar grandes quantidades de informação e gerar a sua representação visual, que pode englobar a construção de tabelas, gráficos ou outro tipo de visualização mais dinâmico, para que os seus utilizadores consigam extrair o máximo de informação possível. Os observatórios têm de ser capazes de associar automaticamente enormes quantidades de informação, sendo isto muitas vezes um processo complexo e demorado de implementar. Como todas as primeiras versões de uma aplicação, ferramenta ou mesmo de um *Website*, existem sempre alguns erros que apenas após vários testes, ou quando usados em situações reais, é que são identificados. Durante o presente trabalho, houve sempre o compromisso de tentar corrigir os erros identificados.

Com o estudo efectuado sobre a visualização de informação e com a análise de observatórios de natureza idêntica ao trabalhado nesta dissertação, foi possível comprovar que existe uma grande dificuldade em representar visualmente grandes quantidades de informação, principalmente quando esta se encontra associada a várias dimensões. O problema acentua-se quando os dados são espaço-temporais e se o pretendido é visualizá-los de uma forma que permita reconhecer padrões ou tendências. Apesar de existir um variado leque de formas, métodos ou representações visuais, nem todos se adequam aos dados a representar e muitos deles são extremamente complexos para o utilizador comum entender. Um dos objectivos deste trabalho era resolver o problema da visualização que o Observatório possuía, principalmente após se ter cumprido o objectivo da implementação da possibilidade de criar indicadores com quatro dimensões. O Observatório tinha que ser capaz de representar dados até quatro dimensões, oferecendo também uma visualização eficaz para a

dimensão espaço-temporal. O modo implementado nesta dissertação para resolver este problema foi baseado no uso de técnicas simples, como as tabelas e gráficos de barras, mas sem mostrar esta informação toda de uma vez só, fornecendo a interactividade necessária ao utilizador para escolher o que deseja observar. Foi também estudada a visualização em mapas e revistos vários exemplos para demonstrar esta informação. Com isto, foi possível melhorar a forma de visualização inicialmente presente no Observatório, que continha várias lacunas.

Foi possível verificar que existem muitos factores que são necessários ter em conta na visualização. Cores, Tamanhos, Formas, Posições e Texturas são apenas alguns que podem tornar as visualizações muito mais intuitivas, atractivas e passíveis de transmitir relações e tendências.

Na procura de soluções para o problema da visualização foi possível constatar que existe uma escassez de ferramentas de visualização, capazes de permitir a construção de representações adequadas para Dados de natureza complexa, permitindo-se, assim, a construção do Conhecimento.

O segundo grande tema estudado nesta dissertação foi a extracção de padrões de utilização. Existe actualmente uma crescente necessidade de analisar como os utilizadores interagem com os *Websites* de modo a melhorar as suas estruturas e os seus conteúdos. Neste contexto, surgem as técnicas *Data Mining*. Para se aplicar este tipo de técnicas é necessário ter pleno conhecimento sobre os dados. As análises podem ser feitas a nível de conteúdo, estrutura, utilização do *Website* e perfis de utilizadores. Um dos objectivos desta dissertação consistia na extracção de padrões de utilização no Observatório. Para tal, foi necessário recorrer a uma ferramenta capaz de efectuar variados registos sobre os utilizadores. Constatou-se que existem várias possibilidades de ferramentas e aplicações que permitem este tipo de análise. As análises dos registos na grande maioria das vezes, irão exigir um pré-processamento dos dados de forma a satisfazer verdadeiramente as necessidades dos utilizadores. Este pré-processamento pode tornar-se complicado, dependendo do que os utilizadores pretendam. A visualização destes dados pode também transformar-se num processo complicado, caso se deseje criar representações com os dados dos registos.

Considera-se que a da gestão de utilizadores implementada no Observatório, no âmbito desta dissertação, pode vir a ser um importante passo para todo o processo de criação de indicadores, que engloba a inserção de dados. Esta inserção, usualmente feita por pessoas ligadas à administração do Observatório, pode agora vir a ser efectuada por pessoas que não estejam ligadas directamente ao Observatório mas que possuam os dados a serem inseridos.

No final desta dissertação é possível concluir que os objectivos inicialmente propostos foram atingidos, podendo no entanto haver espaço para melhorias, sobretudo no que diz respeito à visualização dos percursos dos utilizadores.

5.2 - Futuros Desenvolvimentos

Durante o desenvolvimento desta dissertação foram identificados alguns assuntos que justificam um estudo mais aprofundado do seu comportamento, de modo a melhorar o desempenho do Observatório Social.

Um desses assuntos passa pelo estudo da forma como os utilizadores interagem com as novas visualizações implementadas e se estas conseguem contribuir para um Conhecimento. Como nesta tese não foi possível observar o comportamento dos utilizadores, crê-se que este passo poderia ser extremamente valioso para futuras implementações, modificações e melhoramentos.

Outro dos assuntos identificados que merece um estudo mais aprofundado é o capítulo dos padrões de utilização. Apesar de terem sido implementadas algumas visualizações, poderá ser importante a implementação de mais relações, ligações ou outros tipos de visualizações. Tal como foi explanado anteriormente, a visualização em grafos implementada será um aspecto passível de futuros desenvolvimentos, pois considera-se que esta poderia ser abordada de um modo mais profícuo para o Observatório.

Será interessante também estabelecer relações entre percursos de utilização e a rede de utilizadores registados no Observatório

Referências

[1] M. Worboys, 1995, GIS, "A Computing Perspective", London, Taylor&Francis.

[2] M. Worboys., 1994, "A Unified Model for Spatial and Temporal Information", The Computer Journal, 37:26-34.

[3] N. Andrienko, G. Andrienko, P. Gatalsky, 2003, "Exploratory Spatio-Temporal Visualization: An Analytical Review", Journal of Visual Languages and Computing, special issue on Visual Data Mining, 14:503-541.

[4] M. P. Armstrong, 1988, "Temporality in spatial databases", *Proceedings GIS/LIS'88*, 2: 880-889.

[5] G. Langran, 1993, "*Time in Geographic Information Systems"*, London, Taylor&Francis.

[6] D. J. Peuquet, N. Duan, 1995, "An Event-Based Spatiotemporal Data Model (ESTDM) for Temporal Analysis of Geographical Data", International Journal of Geographical Information Systems 9(1):7-24.

[7] M Yuan, 1999, "Representing Geographic information to enhance GIS support for complex spatiotemporal queries", Transactions in GIS Vol. 3(2):137-160.

[8] M. Jern, J. Franzén, 2006, "GeoAnalytics – Exploring spatio-temporal and multivariate data", Linkoping University, Sweden.

[9] P. Gatalsky, N. Andrienko, G. Andrienko, 2004, "Interactive analysis of event data using space-time cube". 8th international conference on information visualisation (IV'04), IEEE Computer Society 145-152.

[10] P. Compieta, S. Di Martino, M. Bertolotto, F. Ferrucci, T. Kechadi, 2007, "Exploratory spatio-temporal data mining and visualization", J. Vis. Lang, Comput., 18:255-279.

[11] R. H. Guting, M. H. Bohlen, M. Erwig, C. S. Jensen, N. A. Lorentzos, M. Schneider, M. Vazirgiannis, 2000, "A foundation for representing and querying moving objects", ACM Transactions on Database System, 25(1):1-42.

[12] N. Pelekis, B. Theodoulidis, I. Kopanakis, Y. Theodoridis, 2004, "Literature Review of Spatio-Temporal Database Models", The Knowledge Engineering Review, 19:235-274.

[13] J. J. Thomas, K. A. Cook, 2005, "Illuminating the Path: The Research and Development Agenda for Visual Analytics", IEEE-PRESS.

[14] W. Aigner, S. Miksch, W. Muller, H. Schumann, C. Tominski, 2007, "Visualizing Time-Oriented Data – A Systematic View", Computers and Graphics, 31:401–409.

[15] S. K. Card, J. D. Mackinlay, B. Shneiderman, 1999, "Readings in Information Visualization - Using Visualization to Think", Morgan Kaufmann Publhisers Inc.

[16] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere and K. L. Ma, 2009 , "Data, Information and Knowledge in Visualization", *IEEE Computer Graphics and Applications* 29(1):12–19.

[17] B. Shneiderman, 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", IEEE Computer Society Press: Washington, 336 – 343.

[18] J. Mignot, 2005, "Helping Engineers and Scientists Avoid PowerPoint Phluff", *IEEE Aerospace Conference*, IEEE, New York, NY.

[19] D. A. Keim, 2001, "Visual Exploration of Large Databases", Communications of the ACM, Vol. 44(8):38–44.

[20] D. A. Keim, 2002, "Information Visualization and Visual Data Mining", IEEE Transactions on Visualization and Computer Graphics, Vol. 7.
[21] C. Tominski, P. Schulze-Wollgast, H. Schumann, 2005, "3D Information Visualization for Time Dependent Data on Maps", Proceedings of International Conference on Information Visualisation, London, UK.
[22] I. Herman, G. Melançon, M. S. Marshall, 2000, "Graph Visualization and Navigation in Information Visualization: A Survey", IEEE Transactions on Visualization and Computer Graphics, 6(1):24-42.
[23] S. K. Card, J. D. Mackinlay, B. Shneiderman, 1999, "Readings in Information Visualization - Using Visualization to Think", Morgan Kaufmann Publishers Inc.
[24] E. Tufte, 1983, "The Visual Display of Quantitative Information", Graphics Press, Cheshire, Connecticut.
[25] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, 2000, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Department of Computer Science and Engineering University of Minnesota.
[26] R. Cooley, P. N. Tan, J. Srivastava, 1999, "Discovery of Interesting Usage Patterns From Web Data", TR 99-022, University of Minnesota.
[27] Y. H. Cho, J. K. Kim, 2004, "Application of Web Usage Mining and Product Taxonomy to Collaborative Recommendations in e-commerce", Expert Systems with Applications 26:233-246
[28] D. VanderMeer, K. Dutta, A. Datta, "Enabling Scalable Online Personalization on the Web", 2000, Proceedings of the 2nd ACM E-Commerce Conference (EC_00), ACM Press, 185–196.
[29] F. Toolan, N. Kushmerick, "Mining Web Logs for Personalized Site Maps", 2002, International Conference on Web Information Systems Engineering.
[30] Y. H. Wu, A. L. P. Chen, 2002, "Prediction of Web Page Accesses by Proxy Server Log", World Wide Web 5(1):67–88.
[31] R. Cooley, 2003, "The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns", ACM Transactions on Internet Technology, 3(2):93-116.
[32] C. Freitas, O. Chubachi, P. Luzzardi, R. Cava, 2001, "Introdução à Visualização de Informações", RITA, Volume VIII, Número 2.
[33] J.J. Garret, 2005, "Ajax: A New Approach to Web Applications", 2005. Disponível em http://www.adaptivepath.com/publications/essays/archives/000385.php.
[34] G. Andrienko, N. Andrienko, 1999, "Interactive Maps for Visual Data Exploration", Internal Journal Geographical Information Science 13(4):355-374.
[35] C. A. Brewer, L. Pickle, 2002, "Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series", Annals of the Association of American Geographers, 92(4):662-681.
[36] R. G. Cromley, E. K. Cromley, 2009, "Chloropleth map legend design for visualizing community health disparities", International Journal of Health Geographics.
[37] Inf@Vis – The Catastrophe of the Space Shuttle - Disponível em http://www.infovis.net/printMag.php?lang=2&num=114.
[38] Observatório Social de Santa Maria da Feira. Disponível em http://obsocial.idit.up.pt
[39] Página pessoal de Manuel Lima. Disponível em http://www.mslima.com/.
[40] SIG – Análise de ocorrência de crimes. Disponível em http://www.geovista.psu.edu/DCcrimeViz/app/.
[41] SIG – Análise de tráfego automóvel. Disponível em http://www.dot.state.fl.us/planning/statistics/trafficdata/.
[42] SIG – Registo de acidentes de automóveis. Disponível em http://www.unm.edu/~dgrint/lvm.html.
[43] SIG – Análises geológicas e biológicas. Disponível em http://www.usgs.gov/.
[44] GoogleEarth. Disponível em http://earth.google.com/.
[45] Java3D. Disponível em http://java.sun.com/products/java-media/3D/.
[46] phpTrafficA – Disponível em http://soft.zoneo.net/phpTrafficA/index.php
[47] Visualizing Economics - http://www.visualizingeconomics.com/.
[48] Pordata - http://www.pordata.pt/.
[49] CommonGIS. Disponível em http://www.iais.fraunhofer.de/index.php?id=1863&L=1/.

[50] GeoVISTA studio. Disponível em http://www.iais.fraunhofer.de/index.php?id=1863&L=1/.
[51] W3C Characterization Activity. Disponível em http://www.w3.org/WCA.
[52] GoogleMaps. Disponível em http://maps.google.pt/.
[53] jQuery. Disponível em http://jquery.com/.
[54] Wikipedia. Disponível em http://pt.wikipedia.org/.
[55] W3Schools. Disponível em http://www.w3schools.com/..
[56] PHP, Javascript Tutoriais. Disponível em http://www.php.net/.