A EXTRAÇÃO DE DADOS NA ANÁLISE DE REDES

Albertina Ferreira[1], Carlos Caldeira[2] & Fernando Olival[3]

[1]Instituto Politécnico de Santarém, Escola Superior Agrária de Santarém

[2]Universidade de Évora. Departamento de Informática

[3]Universidade de Évora. Departamento de História

RESUMO

Nas bases de dados prosopográficas que envolvem o registo de relações, torna-se fundamental proceder à extração dos dados de modo a que estes sejam corretamente utilizados em análise de redes.

Neste trabalho verificamos que é importante identificar os dados introduzidos incorretamente. Na sequência dessa identificação, mencionamos as metodologias seguidas para determinar algumas dessas situações. Sugerimos também os procedimentos a seguir para colocar os dados num formato adequado à sua integração em *software* de análise de redes.

O repositório de dados que utilizamos tem armazenada informação sobre eventos biográficos e relacionais, sendo o tratamento dos dados fundamental para o estudo das redes de relações entre os diversos atores sociais.

Palavras-chave: Base de dados prosopográfica, análise de redes, extração de dados.

ABSTRACT

In the prosopographical databases involving the register of relationships, it becomes essential to carry out the data extraction so that they are correctly used in network analysis.

In this study we found that it is important to identify the data entered incorrectly. Following this identification we mentioned the methodologies used to determine some of these situations. We also suggest the procedures used to put the data in a format which is suitable for integration in network analysis software.

The data repository we use has stored information about biographical and relational events, given that the treatment of data is essential to the study of relationship networks among the various social actors.

Key words: Prosopographical database, network analysis, data extraction.

INTRODUÇÃO

O estudo da teoria de redes no âmbito das ciências físicas e sociais tem sido uma área pela qual os investigadores apresentam grande interesse. Newman *et al*. (2006) comentam que as redes estão em toda parte e que problemas dinâmicos estão na vanguarda da pesquisa em rede, onde há muitas questões ainda sem resposta. Posteriormente Lazer *et al*. (2009) referem que vivemos a vida em rede. No mesmo ano Borgatti *et al.* (2009) reforçam esta ideia ao referirem que a teoria das redes tem possibilitado explicações para os mais diversos fenómenos sociais numa ampla variedade de contextos.

Para Snijders *et al.* (2010), a evolução nas redes sociais é um domínio de investigação com alguma complexidade. Como é que uma rede social evolui? Podemos encontrar leis e derivar modelos que explicam a sua evolução? Como é que as comunidades surgem numa rede social?

Embora os autores anteriormente focados considerem essencialmente redes a funcionar na atualidade, grande parte dos estudos que realizam poderão ser estendidos a outras épocas, bem como a outras sociedades.

O objetivo deste trabalho é a automatização da extração dos dados, a partir da base de dados *SPARES* (Sistema Prosopográfico de Análise de Relações e Eventos Sociais), para

um formato que possa ser interpretado pelo *software* de redes. No decurso dessa extração foram identificados e corrigidos dados que tinham sido introduzidos de modo incorreto, por várias razões.

Este estudo enquadra-se numa das tarefas propostas - *Developing SPARES: social network analysis* - do projeto aprovado e financiado pela FCT[1]: PTDC/HIS-HIS/118227/2010 – Grupos intermédios em Portugal e no Império Português: as familiaturas do Santo Ofício (c. 1570-1773) – Instituição sede: CIDEHUS [2]

METODOLOGIA

Na realização deste trabalho são utilizados os dados disponíveis na base de dados *SPARES*. Trata-se de uma base de dados relacional desenvolvida de acordo com a Ecologia dos Dados (Caldeira, 2011) e construída no sistema de gestão de base de dados relacional *MySQL*. A base de dados está alojada num servidor central com sistema operativo Linux. Pode ser acedida por *ODBC* (*Open Database Connectivity*) e utilizada por diversos clientes, como os sistemas Windows, Linux ou MacOS, entre outros.

A base de dados *SPARES* tem uma natureza prosopográfica, pois tem armazenada informação sobre indivíduos. Considera-se que estes são parte relevante na dinâmica social. Os dados a utilizar encontram-se distribuídos por três séculos (XVI a XVIII), recaindo este estudo sobre aproximadamente 113000 registos. Esta base de dados foi desenvolvida no âmbito do projeto FCOMP-01-0124-FEDER-007360 – Inquirir da Honra: Comissários do Santo Oficio e das Ordens Militares em Portugal (1570 – 1773).

Na **Figura 1** visualiza-se o modelo de dados que suporta a base de dados SPARES.

[1] Fundação para a Ciência e a Tecnologia

[2] Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora.

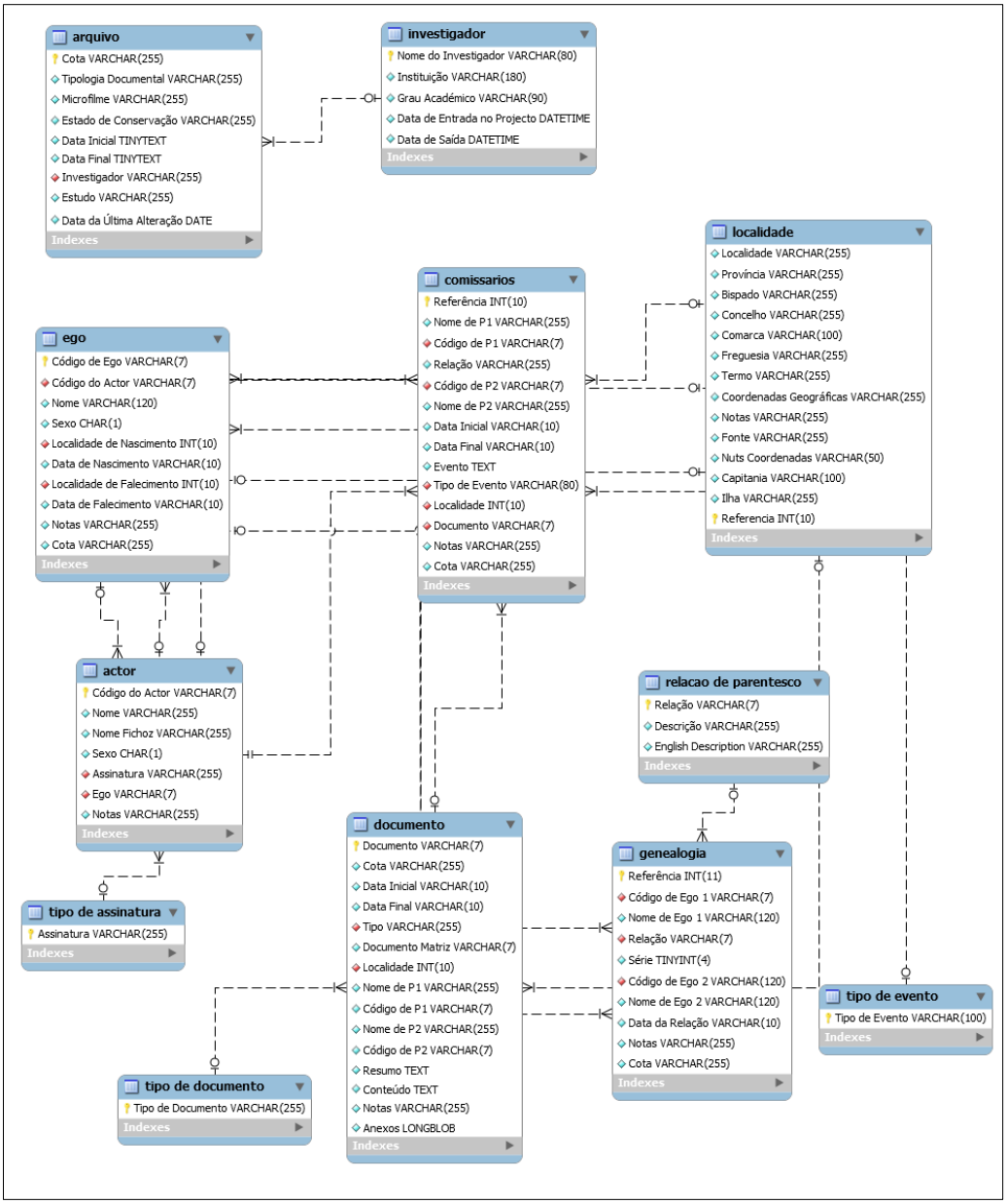


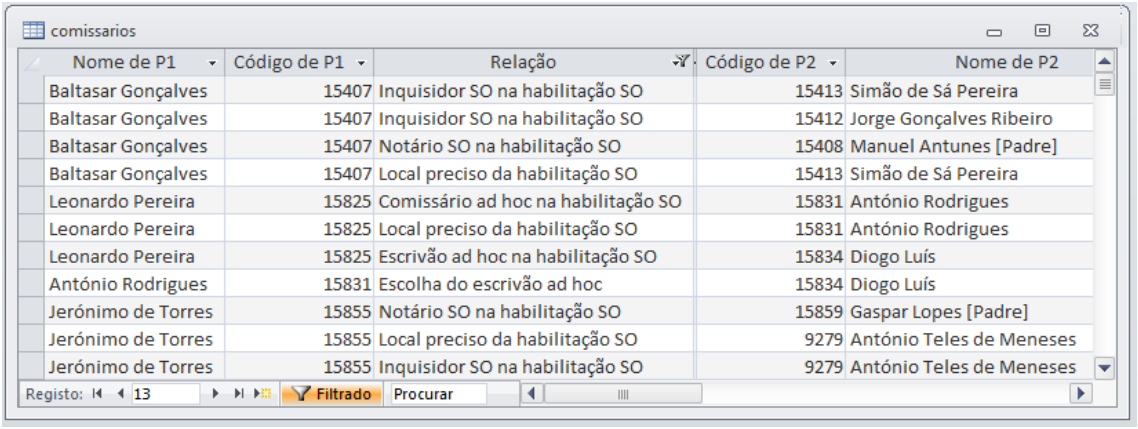
Figura 1 – Modelo de dados do sistema SPARES.

A extração dos dados foi feita considerando a possibilidade de estes serem manipulados por dois *softwares* de rede distintos: *PAJEK* e *GEPHI*.

O *PAJEK* consegue, por um lado, explorar e manipular redes de grande dimensão e, por outro, encontrar-se disponível gratuitamente, para uso não comercial. Pode ser acedido a partir de: http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (Batagelj e Mrvar, 2010; Nooy *et al.*, 2005). Embora através desta aplicação se consigam analisar redes e obter dados, tanto analíticos como gráficos, que podem ser explorados por outras aplicações, pesquisaram-se outras ferramentas *Open Source*. Como futuramente se pretende integrar na mesma plataforma a extração dos dados e a análise de rede,

considerou-se que o *GEPHI* (disponível em: http://gephi.org/) (Bastian *et al.*, 2009) permitirá uma maior interoperabilidade.

O ponto de partida para a extração dos dados é a tabela que se pode observar parcialmente na **Figura 2**:



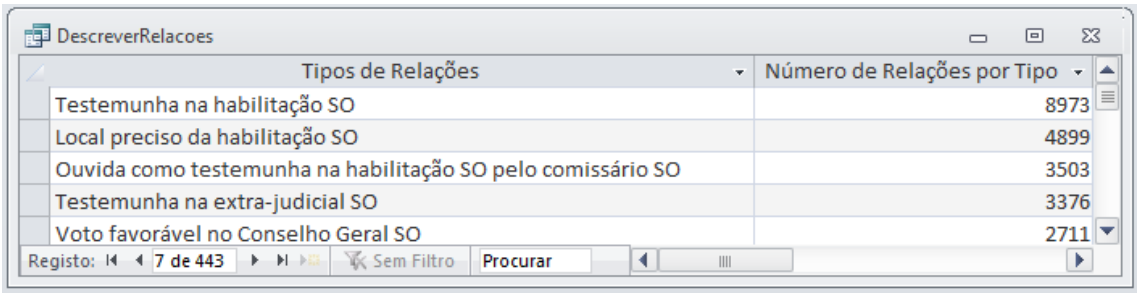
comissarios

Nome de P1	Código de P1	Relação	Código de P2	Nome de P2
Baltasar Gonçalves	15407	Inquisidor SO na habilitação SO	15413	Simão de Sá Pereira
Baltasar Gonçalves	15407	Inquisidor SO na habilitação SO	15412	Jorge Gonçalves Ribeiro
Baltasar Gonçalves	15407	Notário SO na habilitação SO	15408	Manuel Antunes [Padre]
Baltasar Gonçalves	15407	Local preciso da habilitação SO	15413	Simão de Sá Pereira
Leonardo Pereira	15825	Comissário ad hoc na habilitação SO	15831	António Rodrigues
Leonardo Pereira	15825	Local preciso da habilitação SO	15831	António Rodrigues
Leonardo Pereira	15825	Escrivão ad hoc na habilitação SO	15834	Diogo Luís
António Rodrigues	15831	Escolha do escrivão ad hoc	15834	Diogo Luís
Jerónimo de Torres	15855	Notário SO na habilitação SO	15859	Gaspar Lopes [Padre]
Jerónimo de Torres	15855	Local preciso da habilitação SO	9279	António Teles de Meneses
Jerónimo de Torres	15855	Inquisidor SO na habilitação SO	9279	António Teles de Meneses

Registo: 13 Filtrado Procurar

Figura 2 – Dados da base de dados *SPARES*.

Nesta tabela destacamos o atributo Relação. Este será fundamental em futuras análises de redes. Como se pode observar na **Figura 3** existem atualmente 443 relações diferentes, das quais se destaca a "Testemunha na habilitação [do] S[anto] O[fício]" como aquela que possui um maior número de ocorrências.



DescreverRelacoes

Tipos de Relações	Número de Relações por Tipo
Testemunha na habilitação SO	8973
Local preciso da habilitação SO	4899
Ouvida como testemunha na habilitação SO pelo comissário SO	3503
Testemunha na extra-judicial SO	3376
Voto favorável no Conselho Geral SO	2711

Registo: 7 de 443 Sem Filtro Procurar

Figura 3 – Número de relações por tipo.

Quando se iniciou este trabalho, existiam aproximadamente 580 relações. O diferencial que agora apresentamos resulta da correção dos dados que foram identificados como introduzidos incorretamente.

Para preparar os dados de modo a poderem ser utilizados no *software* de rede, foi necessário:

- Criar uma tabela com os códigos e nomes dos primeiros intervenientes (P1);
- Acrescentar a essa tabela os códigos e nomes dos segundos intervenientes (P2);

— Criar tabela com os vértices da rede;

— Criar tabela com as relações da rede;

— Gerar os ficheiros de output que irão ser utilizados na análise da rede.

Para que o ficheiro obtido pudesse ter o formato que o *PAJEK* lê, foi ainda necessário:

— Criar procedimento e pesquisa para atribuir uma numeração sequencial;

— Criar procedimento para eliminar linhas em branco do ficheiro de output.

Uma das análises de rede que se pretende realizar, é obtida por intervalo de tempo. É assim necessário preparar os ficheiros com a informação da década a que cada uma das relações corresponde. A década é determinada tendo como ponto de partida o atributo data, cujo formato é texto. A data pode apresentar-se de duas formas distintas:

— Exatamente esta data, por exemplo 1709=11=08;

— Pensa-se que tenha ocorrido antes desta data, por exemplo 1742<06<09.

Os historiadores precisam de trabalhar deste modo, pois nem sempre têm a certeza da cronologia exata da ocorrência.

No decorrer do trabalho identificaram-se dados introduzidos incorretamente, os quais foram corrigidos, nomeadamente:

— Datas negativas, em matéria de idades, por exemplo;

— Datas anteriores a 1579 (primeira relação conhecida);

— Comissários que mantinham relação com eles próprios;

— O mesmo código (único para cada um dos indivíduos) atribuído a dois indivíduos diferentes;

— O mesmo indivíduo com nomes diferentes, mas com o mesmo código.

A identificação destas ocorrências foi feita através de pesquisas quando se identificou que o ficheiro final possuía mais relações do que as originais. Relativamente às quatro primeiras situações, foram corrigidas manualmente, pois é necessário conhecer o contexto dos dados, nomeadamente as relações envolvidas.

RESULTADOS E DISCUSSÃO

Extração de dados na base de dados SPARES

Apresentamos nas **Figuras 4 e 5** exemplos dos ficheiros obtidos por extração à base de dados *SPARES*. Estes permitirão futuramente a análise de rede nos *softwares* de redes anteriormente apontados. Este estudo foi realizado para uma relação de "Patrocínio".

```
*Vertices    13
    1    "António Correia Bethencourt"    box ic Red bc Red
    2    "António de Noronha e Meneses [Dom]"    box ic Red bc Red
    3    "António Mouzinho [Doutor]"    ic Blue bc Blue
    4    "Bartolomeu César de Andrade"    box ic Red bc Red
    5    "Bento Pais do Amaral"    triangle ic Green bc Green
    6    "Cristóvão de Sousa e Lira [Licenciado]"    box ic Red bc Red
    7    "Diogo Fernandes Branco"    box ic Red bc Red
    8    "Jacome Esteves Nogueira"    ic Blue bc Blue
    9    "João Pais do Amaral"    ic Blue bc Blue
    10    "José de Sousa Castelo Branco [Dom]"    ic Blue bc Blue
    11    "Mariana Isabel de Mesquita e Noronha [Dona]"    box ic Red bc Red
    12    "Martim Filter"    ic Blue bc Blue
    13    "Mateus da Silva"    box ic Red bc Red
*arcs
    10    1
    5    2
    10    4
    9    5
    8    5
    10    6
    12    7
    5    11
    3    13
```

Figura 4 – *Input* para *PAJEK*.

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf>
<graph mode="static" defaultedgetype="directed">
    <nodes>
        <node id="91" label="Bento Pais do Amaral" />
        <node id="151" label="Bartolomeu César de Andrade" />
        <node id="152" label="Cristóvão de Sousa e Lira [Licenciado]" />
        <node id="153" label="José de Sousa Castelo Branco [Dom]" />
        <node id="154" label="António Correia Bethencourt" />
        <node id="2035" label="António de Noronha e Meneses [Dom]" />
        <node id="2351" label="Mariana Isabel de Mesquita e Noronha [Dona]"
        <node id="2375" label="Jacome Esteves Nogueira" />
        <node id="2613" label="Diogo Fernandes Branco" />
        <node id="4731" label="Martim Filter" />
        <node id="6076" label="Mateus da Silva" />
        <node id="6094" label="António Mouzinho [Doutor]" />
    </nodes>
    <edges>
        <edge id="91" source="91" target="2035" label="Patrocínio" />
        <edge id="91" source="91" target="2351" label="Patrocínio" />
        <edge id="153" source="153" target="151" label="Patrocínio" />
        <edge id="153" source="153" target="152" label="Patrocínio" />
        <edge id="153" source="153" target="154" label="Patrocínio" />
        <edge id="2375" source="2375" target="91" label="Patrocínio" />
        <edge id="4731" source="4731" target="2613" label="Patrocínio" />
        <edge id="6094" source="6094" target="6076" label="Patrocínio" />
    </edges>
</graph>
</gexf>
```

Figura 5 – *Input* para *GEPHI*.

Procedimento e pesquisa para numeração sequencial

Para obter o *input* para o *PAJEK*, foi necessário criar um procedimento (**Figura 6**) e uma pesquisa (**Figura 7**), os quais permitissem a atribuição de uma numeração sequencial para os vértices da rede, uma vez que só assim será possível a interpretação pelo *PAJEK* do *input* criado.

```
Public Function numerarSQL(nR) As Long
Static contador As Long

If IsNull(nR) Then
contador = 0
Exit Function
Else
contador = contador + 1
numerarSQL = contador
End If
End Function
```

Figura 6 – Procedimento para numeração sequencial.

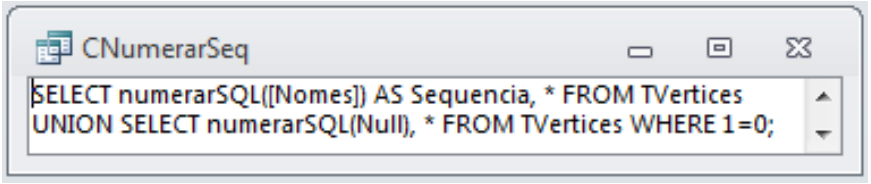


Figura 7 – Pesquisa para numeração sequencial.

Procedimento para eliminar linhas em branco

Após a exportação dos dados para o ficheiro de output, torna-se necessário garantir que este ficheiro não possui linhas em branco (o que normalmente acontece). Criou-se, assim, um procedimento que gerasse um novo ficheiro em que tal não acontecesse. Apresenta-se esse procedimento na **Figura 8**.

```
Dim x As String
Close
Open (CurrentProject.Path & "\RedeEsp.net") For Input As #1
Open (CurrentProject.Path & "\Rede.net") For Output As #2
Do While Not EOF(1)
Line Input #1, x
If x <> "" Then Print #2, x
Loop
Close
```

Figura 8 – Procedimento para eliminar linhas em branco.

Determinação das décadas

As instruções necessárias para que fossem alocadas a cada vértice as décadas em que cada indivíduo interveio nos processos de familiaturas do Santo Ofício, podem consultar-se na **Figura 9**. O intervalo de tempo resultante para cada vértice deverá ter o formato [década x-década y] ou, por exemplo, [7-8], que significa que o indivíduo interveio durante 20 anos, ou seja, 2 décadas.

```
SELECT [Nome de P1], PrimData, SegData, IIf([Primdata]>="1570*" And [PrimData]<"1580*","1",IIf([Primdata]>="1580*" And [PrimData]<"1590*","2",IIf([Primdata]>="1590*" And
[PrimData]<"1600*","3",IIf([Primdata]>="1600*" And [PrimData]<"1610*","4",IIf([Primdata]>="1610*" And [PrimData]<"1620*","5",IIf([Primdata]>="1620*" And
[PrimData]<"1630*","6",IIf([Primdata]>="1630*" And [PrimData]<"1640*","7",IIf([Primdata]>="1640*" And [PrimData]<"1650*","8",IIf([Primdata]>="1650*" And
[PrimData]<"1660*","9",IIf([Primdata]>="1660*" And [PrimData]<"1670*","10",IIf([Primdata]>="1670*" And [PrimData]<"1680*","11",IIf([Primdata]>="1680*" And
[PrimData]<"1690*","12",IIf([Primdata]>="1690*" And [PrimData]<"1700*","13")))))))))))) AS Parte1Prim, IIf([Primdata]>="1700*" And [PrimData]<"1710*","14",IIf([Primdata]>="1710*"
And [PrimData]<"1720*","15",IIf([Primdata]>="1720*" And [PrimData]<"1730*","16",IIf([Primdata]>="1730*" And [PrimData]<"1740*","17",IIf([Primdata]>="1740*" And
[PrimData]<"1750*","18",IIf([Primdata]>="1750*" And [PrimData]<"1760*","19",IIf([Primdata]>="1760*" And [PrimData]<"1770*","20",IIf([Primdata]>="1770*" And
[PrimData]<"1780*","21",IIf([Primdata]>="1780*" And [PrimData]<"1790*","22",IIf([Primdata]>="1790*" And [PrimData]<"1800*","23")))))))))) AS Parte2Prim, IIf([SegData]>="1570*" And
[SegData]<"1580*","1",IIf([SegData]>="1580*" And [SegData]<"1590*","2",IIf([SegData]>="1590*" And [SegData]<"1600*","3",IIf([SegData]>="1600*" And
[SegData]<"1610*","4",IIf([SegData]>="1610*" And [SegData]<"1620*","5",IIf([SegData]>="1620*" And [SegData]<"1630*","6",IIf([SegData]>="1630*" And
[SegData]<"1640*","7",IIf([SegData]>="1640*" And [SegData]<"1650*","8",IIf([SegData]>="1650*" And [SegData]<"1660*","9",IIf([SegData]>="1660*" And
[SegData]<"1670*","10",IIf([SegData]>="1670*" And [SegData]<"1680*","11",IIf([SegData]>="1680*" And [SegData]<"1690*","12",IIf([SegData]>="1690*" And
[SegData]<"1700*","13")))))))))))) AS Parte1Seg, IIf([SegData]>="1700*" And [SegData]<"1710*","14",IIf([SegData]>="1710*" And [SegData]<"1720*","15",IIf([SegData]>="1720*" And
[SegData]<"1730*","16",IIf([SegData]>="1730*" And [SegData]<"1740*","17",IIf([SegData]>="1740*" And [SegData]<"1750*","18",IIf([SegData]>="1750*" And
[SegData]<"1760*","19",IIf([SegData]>="1760*" And [SegData]<"1770*","20",IIf([SegData]>="1770*" And [SegData]<"1780*","21",IIf([SegData]>="1780*" And
[SegData]<"1790*","22",IIf([SegData]>="1790*" And [SegData]<"1800*","23")))))))))) AS Parte2Seg, IIf([parte1Prim]="1" Or [parte1Prim]="2" Or [parte1Prim]="3" Or [parte1Prim]="4" Or
[parte1Prim]="5" Or [parte1Prim]="6" Or [parte1Prim]="7" Or [parte1Prim]="8" Or [parte1Prim]="9" Or [parte1Prim]="10" Or [parte1Prim]="11" Or [parte1Prim]="12" Or
[parte1Prim]="13",[parte1Prim],[Parte2Prim]) AS JuntoPrim, IIf([parte1Seg]="1" Or [parte1Seg]="2" Or [parte1Seg]="3" Or [parte1Seg]="4" Or [parte1Seg]="5" Or [parte1Seg]="6" Or
[parte1Seg]="7" Or [parte1Seg]="8" Or [parte1Seg]="9" Or [parte1Seg]="10" Or [parte1Seg]="11" Or [parte1Seg]="12" Or [parte1Seg]="13",[parte1Seg],[parte2Seg]) AS JuntoSeg,
IIf([JuntoPrim]=[JuntoSeg],"[" & [JuntoPrim] & "]","[" & [JuntoPrim] & " - " & [JuntoSeg] & "]") AS Decadas
FROM CDatasInicialP1
WHERE [Nome de P1] Is Not Null;
```

Figura 9 – Pesquisa para determinação das décadas.

Correção de dados

Como já foi referido, identificaram-se diversas situações de introdução incorreta de dados que, pela sua especificidade, foram corrigidos manualmente.

Na **Figura 10** podem visualizar-se dois registos onde foi detetada a introdução incorreta como resultado da presença de erros ortográficos.

C_ListaRelações

Relação	Nr de ocorrências
Notário SO na extra-judicial SO	14
Notário SO na habiçitação SO	2
Notário SO na habilitação SO	1
Notário SO na habilitação SO	1395

Figura 10 – Introdução incorreta de dados por erro ortográfico.

Identificaram-se diversos códigos P1 e P2 em que os nomes atribuídos eram diferentes, o que não pode acontecer. Na **Figura 11** pode observar-se a pesquisa que foi realizada para identificar todas as situações referentes a P1. Na **Figura 12** apresenta-se o resultado dessa pesquisa. Procedeu-se de modo idêntico para as ocorrências de P2.

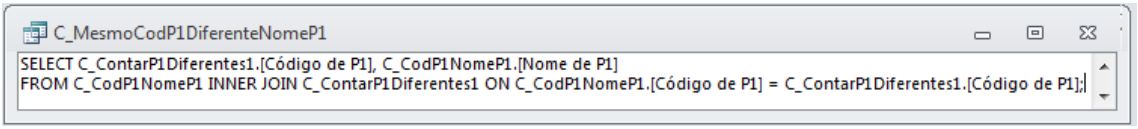


Figura 11 – Pesquisa para identificação da atribuição de nomes diferentes para o mesmo código P1.

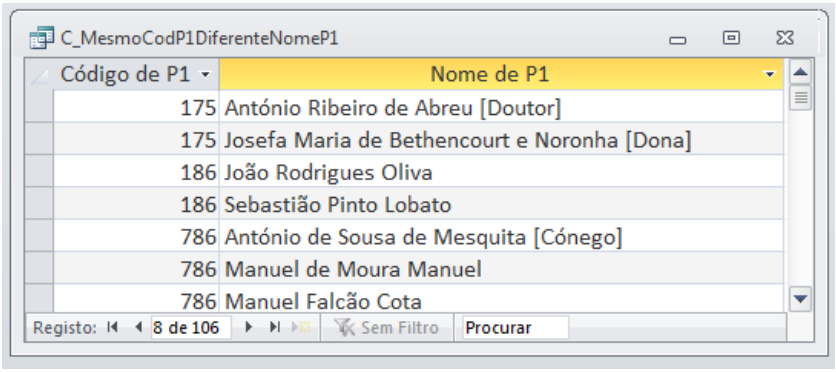


Figura 12 – Identificação da atribuição de nomes diferentes para o mesmo código P1.

CONCLUSÕES E TRABALHO FUTURO

Com este estudo, espera-se ter demonstrado que a extração adequada dos dados é um passo importante para a análise de redes.

No decurso do trabalho, e em estreita colaboração com os membros do projeto, foi ainda possível identificar e corrigir algumas situações resultantes da introdução incorreta de dados. A identificação destas situações e a sua posterior correção é fundamental, pois quem introduz grandes números perde facilmente o controlo dos dados, muitas vezes por distração.

Como desafio futuro, pretende-se construir uma aplicação que permita a adequação entre a base de dados prosopográfica *SPARES* e o *software* de redes *GEPHI*. Deste modo, qualquer utilizador de Ciências Sociais e, como tal, menos familiarizado com a Estatística e a Informática, poderá realizar facilmente uma análise na rede social que estuda.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Newman, M. E. J.; Barabási, A.; Watts, D. J. 2006. The Structure and Dynamics of Networks [em linha]. [Acedido: 3 de Abril de 2013]. Disponível em: http://press.princeton.edu/chapters/s8114.html

[2] Lazer, D.; Pentland, A.; Adamic, L; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; Alstyne, M. V. 2009. Life in the Network: the Coming Age of Computational Social Science. *Science* 323(5915): 721–723. doi: 10.1126/science.1167742.

[3] Borgatti, S. P.; Mehra, A.; Brass, D. J.; Labianca, G. 2009. Network Analysis in the Social Sciences. *Science* 323: 892-895.

[4] Snijders, T.A.B.; Steglich, C.E.G.; van de Bunt, G.G. 2010. Introduction to Actor-Based Models for Network Dynamics. *Social Networks* 32: 44-60.

[5] Caldeira, C. 2011. A Arte das Bases de Dados. Edições Sílabo, Lisboa. ISBN 978-972-618-627-4

[6] Batagelj, V; Mrvar, A. 2010. *Pajek: Program for Analysis and Visualization of Large Networks. Reference Manual List of commands with short explanation version 2.00*. University of Ljubljana. Slovenia.

[7] Nooy, W; Mrvar, A; Batagelj, V. 2005. *Exploratory Network Analysis with Pajek*. Cambridge University Press. New York.

[8] Bastian, M; Heymann, S; Jacomy, M. 2009. Gephi: An open source software for exploring and manipulating networks. *In Proceedings of the Third International ICWSM Conference*. California, USA. 361-362.