Carvalho, Joaquim (1997), "Comportamentos Morais e Estruturas Sociais numa paróquia de Antigo Regime (Soure, 1680-1720)." Dissertação de Doutoramento apresentada à Faculdade de Letras da Universidade de Coimbra, Parte3.

3. MÉTODOS

3.1. AS NECESSIDADES METODOLÓGICAS E TÉCNICAS DA RECONSTITUIÇÃO DE COMUNIDADES HISTÓRICAS

3.1.1. INTRODUÇÃO

Nesta parte abordaremos os aspectos metodológicos e técnicos subjacentes aos conteúdos apresentados nas partes anteriores. Foi uma preocupação central desta investigação criar instrumentos e processos reaproveitáveis que rentabilizassem o laborioso trabalho de reconstituição de comunidades históricas. Esse objectivo baseava-se na convicção profunda que o tipo de trabalho que aqui se esboçou, alicerçado no cruzamento da informação nominal dispersa em várias fontes, é essencial ao esclarecimento da lógica interna de determinados processos históricos, e só não é uma abordagem mais

frequentemente seguida devido às dificuldades técnicas de gerir de forma eficaz uma grande quantidade de informação multi-variada.

Grande parte do investimento que está por detrás desta investigação foi dirigido para a construção dos "instrumentos" que pudessem facilitar e eventualmente vulgarizar uma abordagem micro-histórica, que neste caso concreto assume a forma de uma "reconstituição de comunidades históricas".

O leitor a quem estas linhas se dirigem não é muito fácil de definir e consequentemente não é muito fácil acertar o vocabulário e o estilo na exposição do que é de facto um labor interdisciplinar. Algumas secções desta parte serão bastante técnicas, do ponto de visto informático. Outras terão características mais genéricas. Todas, contudo, são fruto das preocupações concretas de um historiador e das suas necessidades ao nível de instrumentos de trabalho. Sem compreender essas necessidades e a problemática científica que está por detrás delas não se entenderá a razão de ser das construções técnicas e metodológicas que se seguem.

Assim, para além do raro historiador com forte *background* informático, este capítulo serve dois públicos distintos e, infelizmente, muitas vezes estanques: o dos historiadores que queiram perceber um pouco mais detalhadamente como certo tipo de investigações podem ser informatizadas e o informático que, na posição de exercer o seu ofício num projecto historiográfico, queira saber como determinados problemas foram resolvidos.

Para facilitar uma abordagem parcelar fornecemos um guia dos capítulos que seguem.

Este primeiro capítulo procura, sem entrar em grandes detalhes, enquadrar e definir os objectivos metodológicos que foram perseguidos aqui. Isso será feito a três níveis, sucessivamente mais específicos:

- um enquadramento geral que refere a inspiração teórica por detrás deste tipo de investigação histórica;
- uma definição, esperemos que clara, daquilo que caracteriza a abordagem micro-histórica e em especial a reconstituição de comunidades históricas;
- finalmente, para fazer a ponte para os restantes capítulos, uma enumeração daquilo que se espera de um sistema informático que suporte a reconstituição de comunidades históricas.

O segundo capítulo trata da estrutura de uma base de dados especialmente pensada para a reconstituição de comunidades históricas. São identificados os problemas com que habitualmente se defrontam os historiadores quando tentam usar programas de bases de dados com este tipo de dados. Propõe-se uma estrutura genérica capaz de resolver esses problemas. Descreve-se em detalhe o modo concreto de implementação dessa estrutura.

Este capítulo destina-se sobretudo a quem tenha a seu cargo o desenho e manutenção de bases de dados prosopográficas. Interessa a historiadores que pretendam ter uma visão das questões informáticas envolvidas neste tipo de estudos. Alguns formalismos descritos poderão ter interesse genérico para quem esteja envolvido em modelização de dados, como por exemplo o modo como se integrou o modelo orientado a objectos com o modelo relacional. Embora não inclua nada de transcendente, muitas secções deste capítulo revestem um aspecto bastante técnico.

O terceiro capítulo explica a forma como se processa a transcrição de fontes para a estrutura anteriormente descrita. Descreve-se uma linguagem de registo de dados pensada para facilitar a transcrição de fontes com percas mínimas de

informação. Explica também como são traduzidas as fontes transcritas nessa linguagem e como o resultado da tradução é inserido na estrutura da base de dados. Este capítulo é útil para os historiadores que se queiram familiarizar com a metodologia de transcrição de fontes que aqui foi usada. Do ponto de vista mais técnico descreve o funcionamento dos tradutores, programas informáticos que procedem à conversão de documentos próximos da fonte nos moldes de uma estrutura de base de dados relacional.

O quarto capítulo descreve o processo de cruzamento nominal, ou seja a forma como as várias referências a pessoas, dispersas pelas várias fontes, são consolidadas em biografias. É porventura o capítulo mais complexo, ao descrever os procedimentos utilizados e a estatística por detrás do processo.

Estes capítulos, apesar de por vezes entrarem em detalhes técnicos, não explicitam os detalhes concretos de uma implementação informática específica. O seu objectivo é demonstrar os formalismos e algoritmos que estão por detrás deste empreendimento, mas não descem ao detalhe de como implementar concretamente. Os apêndices explicitam alguns elementos considerados mais significativos da implementação feita desses conceitos.

3.1.2. ENQUADRAMENTO GERAL

A justificação epistemológica da importância daquilo que chamamos, para simplificar, de perspectiva micro-histórica, já foi suficientemente feita. Um texto de referência, com título feliz, é o de Ginsburg e Poni, "il nome e il

comme" (o nome e o como)[1]. Preferimos, por nos parecer mais adequado ao espírito do investigador, chamar "abordagem nominalista" à tentativa de recolocar o indivíduo no centro metodológico da pesquisa, fazendo do labor de atingir o "como" (a explicação) essencialmente uma tarefa de seguir o "nome" (os indíviduos). Não se trata de já antigas polémicas sobre humanismo e estruturalismo. Pelo contrário, esta abordagem é quase um pragmatismo científico. Entende que para compreender o que se passou é necessário perceber como os factos concretos foram perspectivados pelos actores históricos. Isso passa pela reconstituição da história de vida de muitas pessoas. O objectivo final é sempre uma explicação coerente e satisfatória do modo de funcionamento de algum aspecto da sociedade passada, aspecto esse que pode ter contornos estruturais, institucionais e até abstractos. A diferença aqui é que os indivíduos concretos envolvidos nesses processos e nessas estruturas tornam-se o objecto principal de focalização do interesse do historiador, convencido que é nos indíviduos que se colhem os traços das grandes lógicas da História e que, inversamente, é pela interacção de pequenas e individuais motivações que se sustentam os grandes movimentos. A "abordagem nominalista", se nos é permitida a tentativa de fazer passar uma etiqueta ligeiramente provocante, é antes do mais uma exigência de inteligibilidade do "trabalhador de campo" e não uma formulação de teórico preocupado com os fundamentos filosóficos da sua ciência.

---

[1] Ginzburg, C.; Poni, C. - *Il nome e il come: scambio ineguale e mercato storiografico*. "Quaderni Storici", ano 14, nº1 (Gennaio-aprile 1979), p. 181-190.

3.1.2. O QUE SE ENTENDE POR "RECONSTITUIÇÃO DE COMUNIDADES HISTÓRICAS" (RCH)

A forma pela qual a abordagem "nominalista" ou "micro-histórica" se manifesta neste trabalho é o da "reconstituição de comunidades históricas" (RCH). O conceito de reconstituição de comunidades históricas é antigo na historiografia. Na Grã-Bretanha, por clara inspiração das escolas antropológicas, o conceito foi assumido como programa metodológico[2]. Existem exemplos marcantes dessa abordagem de cariz antropológico como Earls Clone e Terling[3] .

Uma reconstituição de uma comunidade histórica é uma tentativa de recolher e cruzar toda a informação sobre as pessoas que habitaram em comunidade um determinado espaço. Em princípio assume-se que a comunidade sob observação tem uma certa coerência funcional, isto é, encerra em si mesma os principais mecanismos que permitem a sua auto-preservação e que, em consequência, a sua análise detalhada revelará a informação necessária para compreender o seu funcionamento interno. Os exemplos mais comuns, para o antigo regime, são a paróquia ou o município, espaços que organizam funcionalmente a vida dos indivíduos e que operam com relativa autonomia, pelo menos nas mais básicas tarefas do quotidiano (reprodução, produção de riqueza, exercício do poder, etc....).

Apesar da acepção espacial do conceito de "comunidade" muitas das técnicas e formalismos aqui descritos aplicam-se a interpretações mais

---

[2] Sobre a influência da Antropologia na História e as consequências dessa interacção no plano informático ver:Rowland, Robert -*L'informatica e il mestiere dello storico*. "Quaderni storici". Vol. 78 (1991), p. 693-720; *Antropologia, história e diferença: alguns aspectos.* Porto: Edições Afrontamento, 1987.

[3] Macfarlane, Alan, *Reconstructing Historical Communities,* Cambridge, Cambridge University Press. 1977. Wrightson, Keith; Levine, David, *Poverty and Piety in an English Village. Terling, 1525-1700*. N.Y. : San Franc. : London: Academic Press, 1979.

alargadas e elásticas do conceito. Podemos centrar análises nominalistas sobre grupos sociais não necessariamente ancorados no espaço mas com uma coerência interna forte, construída sobre afinidades sociais e redes densas de relações recíprocas. Aí a "comunidade" designa um conjunto de pessoas, atributos, funções e relações que funcionam como um todo e como um todo têm de ser compreendidas. É necessário por isso "reconstituir" o grupo na sua complexidade interna para fundamentar uma aproximação inteligível. Nesse sentido o estudo de um grupo social é semelhante à reconstituição de uma paróquia ou de um concelho. A RCH é por isso muito próxima do conceito de prosopografia ou biografia colectiva[4].

Inversamente são possíveis muitas abordagens de cariz globalizante, pelo menos ao nível da quantidade e variedade das fontes utilizadas, que não se enquadram dentro daquilo que aqui se designa por reconstituição de comunidades históricas. O exemplo mais claro será o da "monografia local" clássica, muitas vezes obra de eruditos locais, que funciona como enciclopédia de um micro-cosmos onde as pessoas concretas não são objectos centrais da atenção do investigador, excepto, evidentemente, os notáveis habituais[5].

O mesmo raciocínio se aplica à reconstituição de famílias clássica, oriunda da demografia. Aqui as relações com o conceito de RCH são mais estreitas mas não permitem qualquer confusão. O demógrafo que efectua a reconstituição de paróquias tem questões concretas e utiliza técnicas muito precisas para as

---

4 Millet, Hélène (dir.) -*Informatique et prosopographie: Actes de la Table Ronde du CNRS, Paris, 25-26 Octobre 1984, textes réunis par Hélène Millet* . Paris: ed. CNRS, 1985; Bulst, Neithard, *Prosopography and the computer: problems and possibilities.* In: Denley, Peter; Fogelvik, Stefan; Harvey, Charles (eds), *History and Computing II,* Manchester 1989, p.12-18.

5 Um exemplo da obra de erudito local, preciosa pela quantidade de informação, muita dela inédita e fruto da prolongada permanência in loco é, para Soure, Conceição, Augusto, *Soure*.,Coimbra,1942.

responder. É certo que, de todos os especialistas que se debruçam sobre fontes locais, é ele que de facto centra toda a sua análise sobre as pessoas. Mas não são as pessoas em si que o preocupam. São sim as grandes variáveis da vida de uma população, a mortalidade a nupcialidade e a mortalidade. As pessoas são um passo necessário, porque sem o cruzamento dos registos, sem a reconstituição das famílias, a maior parte das variáveis "interessantes" são inalcançáveis: idade média ao casamento, intervalos intergenésicos, gravidezes pré-nupciais, duração média de vida. Mas as pessoas só interessam para essa ligação restrita entre os registos que faz o tempo dos fenómenos da vida mensurável. Num baptismo, por exemplo, os demógrafos ignoram a informação sobre os padrinhos. Ou as testemunhas nos casamentos. Em ambos os casos os indivíduos não ocorrem no acto numa função demograficamente pertinente. O registo dessa ocorrência não trará nenhuma informação demográfica adicional. Quando muito o demografo registará a ocoorrência pelo valor marginal que tal informação mais tarde poderá ter para resolver uma dúvida ou dificuldade de identificação. Para o "nominalista" a função de "padrinho" é relevante em si, pelo seu significado social e pelo papel que tem na construção das redes de relações interpessoais. Muito daquilo que faz a essência do trabalho de reconstituição de comunidades está fora das preocupações dos demógrafos clássicos[6].

Nada disto constitui uma diminuição da importância crucial dos conceitos e técnicas apuradas pela demografia ao longo dos anos nos métodos envolvidos na reconstituição de comunidades históricas. Muitos dos resultados da demografia paroquial não relevam exclusivamente da problemática dessa

---

[6] Para uma perspectiva recente que defende a necessidade de reconstituição de famílias se aproximar mais da reconstituição de comunidades ver King, Steven - *Historical Demography, Life-cycle Reconstruction and Family Reconstitution: New Perspectives,* "History and Computing", v.8, n.2, p. 62-77.

disciplina. Antes têm consequências centrais sobre a interpretação da sociedade, dos aspectos mentais, da difusão de determinados padrões de comportamento, de novas concepções de vida e de atitudes socialmente demarcadas. São assim centrais. Mas a reconstituição das famílias não é, só por si, uma reconstituição de comunidades, no sentido que aqui tentamos definir.

Poderá parecer que o que aqui se designa por RCH é apenas a tentativa de acumular tudo o que se pode saber sobre todas as pessoas de uma localidade, na esperança de conseguir alguma luz sobre o funcionamento da realidade social. É, basicamente, disso que se trata, sem complexos. Foi necessário anos de observações aparentemente sem objectivo para perceber que a trajectória dos planetas era elíptica e não circular. A importância desse trabalho de observação puro foi crucial na modificação da visão do universo da época. Não se pode compreender o que não se vê. E em muito casos não se pode ver sem o instrumento de observação adequado.

A RCH compõe-se de várias reconstituições sobrepostas.

A recolha e cruzamento da informação nominal permite recuperar biografias individuais que são a base de todo o trabalho posterior. Sobre essas biografias, que reunem toda a população sobre observação, reconstituem-se planos de realidade de complexidade crescente.

Em primeiro lugar a história demográfica dos indivíduos, que corresponde ao nível da tradicional reconstituição de famílias. Daqui extraem-se não só os grandes descritores demográficos da comunidade, as crises, a evolução das variáveis vitais, mas também, ao levar o olhar para o micro-cosmos de cada vida, os acontecimentos relevantes que nos permitem avaliar determinados trajectos: as concepções pré-nupciais, as idades ao casamento, o impacto das crises e das epidemias em biografias individuais.

Acima das famílias e das informações dos actos vitais reconstituem-se as redes mais alargadas de relações inter-pessoais: as genealogias, as alianças cruzadas entre famílias, os papéis sociais do apadrinhamento e da procuração. Logo aqui se identificam indivíduos centrais pela frequência com que aparecem como padrinhos de baptismos ou testemunhas em casamentos. Vemos também como esses papéis de centralidade social são transmitidos de geração em geração de pais para filhos, de sogros para genros, de maridos para viúvas, etc... As linhagens entrelaçam-se com o parentesco espiritual e permitem reconstruir redes de influência, entrelaçadas ciclicamente, mas, surpreendentemente, mantendo por vezes universos de sociabilidade distintos dentro de níveis sociais semelhantes num espaço apesar de tudo muito reduzido geograficamente.

Vemos depois como estas redes de sangue e alianças recobrem as clientelas do poder. Vemos os cargos políticos e administrativos a circularem por essas redes, passando de geração em geração, ou de família a família por alianças, e o mesmo se passa com as propriedades suficientemente importantes para conservar. Cargos e bens materiais têm uma funcionalidade semelhante nas estratégias patrimoniais e circulam do mesmo modo. Assim o quadro institucional, os vários orgãos que concedem capital material ou imaterial, funciona como um espaço em que os actores encenam as suas estratégias. Sem a reconstrução desse espaço, enquanto campo de possibilidades, as acções individuais não são compreensíveis.

A circulação de bens que as escrituras notariais relatam permite reconstruir o fluxo de bens e o modo como as relações económicas se entrelaçam com os outros planos da comunidade.

As transacções relativas a bens imóveis, junto com os tombos, possibilitam a construção de uma imagem muito necessária da propriedade. Saber a quem pertence a terra, desenhar o mapa da propriedade ou uma aproximação ao

mesmo, é um elemento fundamental da reconstituição da comunidade. A produção elementar de bens depende intimamente da organização da propriedade, dos ritmos impostos pelas condições geográficas e climatéricas, dos encargos, limites e rituais que socializam a produção. Aí também se configura um campo de possibilidades que permite situar e tentar compreender as escolhas individuais.

Finalmente na base da RCH está uma apreensão do espaço que representa a infarestrutura mais primária da comunidade. O espaço é simultaneamente uma figura de encerramento e uma possibilidade de relação. Reconstruir o espaço significa saber quem e o quê estava onde e quando. Tarefa tão complexa como a reconstituição das biografias individuais e intimamente entrelaçada com esta. Soure, com a sua extensão invulgar e com uma elevada dispersão do povoamento, não foi um caso fácil de reconstituição do espaço.

Todos estes níveis interligados de um empreendimento de RCH implicam um tratamento intenso de informação muito dispersa. A aplicação de uma metodologia informática a investigações desta natureza é, muitas vezes, uma evidência e um desapontamento. Com efeito, tudo na quantidade e complexidade dos dados envolvidos apontaria para um uso maciço de computadores. Mas precisamente quando se tenta levar a cabo esse desidério as dificuldades encontradas multiplicam-se.

Do ponto de vista da tecnologia actual uma aplicação de RCH apresenta um desafio considerável. Esse desafio só em parte é aquilo que poderia parecer à primeira vista mais evidente: a necessidade de máquinas altamente capazes, altamente velozes, com grande capacidade de armazenamento de informação. Tudo isso é verdade mas não constitui o essencial da questão. O essencial é que a tarefa de RCH põe problemas de formalização, representação dos dados, modelização de processos de tratamento da informação e manuseamento

inteligível de resultados enormemente complexos. O principal objectivo deste capítulo é explicitar as escolhas e soluções tomadas e encontradas nessa área.

Como em todos os problemas tecnológicos não existe aqui uma solução óptima em si. A melhor solução procura-se sempre em função de um cenário de utilização que é definido por determinados objectivos. Não há uma solução ideal e perfeita para a informatização da RCH. O conjunto de soluções, formalismos e metodologias que aqui apresentamos devem ser avaliados como a adequação da tecnologia disponível aos objectivos que enumeraremos a seguir.

3.1.3. OBJECTIVOS PARA UM SISTEMA INFORMATIZADO DE RCH.

O primeiro objectivo de um sistema informatizado de reconstituição de comunidades históricas é dar a primazia à fonte. Entendemos por isto que a fonte original deve ter um lugar central na representação da informação. Os anglo-saxónicos utilizam a expressão "source oriented" para designar um sistema que procura manter o mais intacta possível a informação da fonte original. Os sistemas "source-oriented" opõem-se aos sistemas orientados para os resultados, em que a informação é registada em função dos resultados que se pretendem e sem intenção de preservar a estrutura original[7].

---

[7] Um desenvolvimento tecnológico fundamental para a afirmação de uma abordagem orientada para as fontes foi o desenvolvimento do sistema Kleio por Manfred Thaller, centrado na necessidade de podermos representar os dados oriundos das fontes históricas em computador, salvaguardando toda a sua riqueza: Thaller, Manfred, *The need for a Theory of Historical Computing* In: Denley, Peter; Fogelvik, Stefan; Harvey, Charles (eds), *History and Computing II,* Manchester 1989, p.2-11.; *Kleio: a database system*. Gottingen: Max-Planck-Institut, 1993; *The archive on top of your desk.* "Historical Methods", vol. 28 (Jun 1995), p. 133-152.

A primazia à fonte é uma decisão metodológica importante porque se baseia num objectivo fundamental: a reversibilidade e validação dos resultados. Por reversibilidade entende-se aqui a possibilidade de retraçar os passos dados entre a informação primária e as conclusões finais. Um sistema é reversível se, por exemplo, perante uma extensa biografia reconstituída, existe a possibilidade técnica de rever todas as decisões de identificação que foram tomadas e todas as informações parcelares no seu contexto documental original. Esta característica suporta a validação de resultados porque permite a um investigador independente rever ou refazer as decisões tomadas.

Se uma base de dados só inclui informação que foi selectivamente retirada de várias fontes com o objectivo de esclarecer um determinado fenómeno, então os dados disponíveis já encerram uma série de decisões implícitas que operaram durante a fase da recolha. Essa selecção prévia da informação pertinente marca os dados e não permite usá-los para rever o processo de investigação e eventualmente confirmar a validade das conclusões tiradas.

Um sistema orientado para a preservação da fonte é um sistema que regista o máximo de informação do documento original e opera as escolhas, selecções e tratamentos *a posteriori*, de forma explícita. Como os passos mais significativos são feitos dentro do sistema, podem mais facilmente ser desfeitos e revistos, ou confrontados com procedimentos alternativos.

Basicamente a preservação da fonte utiliza-se seja porque não se pode prever *a priori* o tipo de tratamento que os documentos vão ter, seja porque se pretende preservar dentro do computador a informação original para refazer ou validar os processos que chegaram até às conclusões finais.

Os sistemas orientados para a fonte são mais custosos em termos de tempo, requisitos técnicos e sofisticação formal. É mais trabalhoso registar a fonte do que pré-seleccionar só os elementos que se julgam pertinentes. O registo dos documentos em forma extensa, próxima do original, ocupará mais as

capacidades dos computadores do que informações elementares pré-digeridas. Finalmente, para obter informação relevante do documento transcrito é necessário formalizar procedimentos de registo e tratamento que podem ser bastante complexos.

Esta diferença entre uma abordagem orientada para a fonte e uma abordagem mais selectiva não é exclusiva de ambientes informáticos. A principal diferença entre o método de reconstituição de famílias proposto por Norberta Amorim e o método tradicional de Fleury-Henry reside no facto de o primeiro ser orientado para os resultados e o segundo orientado para as fontes[8]. Norberta Amorim não regista os casamentos, os baptismos e os óbitos. Como o objectivo da reconstituição de famílias é basicamente identificar os casais, a informação dos registos paroquiais é imediatamente registada em fichas de família. A maior parte das decisões de identificação é tomada quando, perante um casal que ocorre na fonte, se decide acrescentar informação a uma ficha existente ou abrir uma ficha nova. Como muitas decisões são tomadas no processo de recolha este não é reversível, isto é, não é muito fácil passar as fichas a um segundo investigador para validar o processo: a maior parte das decisões está embuída nas próprias fichas.

O método tradicional faz uma ficha para cada acto, sem qualquer tipo de decisão no momento da recolha. Só depois de todos os actos recolhidos começa o processo de identificação. Em teoria o mesmo conjunto de fichas pode ser fornecida a investigadores diferentes criando uma situação de controle. As

---

[8] Amorim, Maria Norberta Simas Bettencourt - *Método de exploração dos livros de registos paroquiais e reconstituição de famílias,* Guimarães, Ed.A., 1982; Amorim, Maria Norberta; Lima, Luís - *Demografia Histórica e micro-informátic: uma experiência sobre uma paróquia açoriana.* "Boletim do Instituto Histórico da Ilha Terceira", Angra do Heroísmo, 44, 1986. Fleury, Michel; Henry, Louis - *Nouveau manuel de dépouillement et d'exploitation de l´état civil ancien,* Paris, INED, 3ªed, 1985. Henry, Louis - *Simulation d'une reconstitution de familles par ordinateur*. "Annales Demographie Historique", (1972), p. 303-309.

fichas de baptismos, casamentos e óbitos, não contêm em si próprias grandes decisões sobre quem é quem. São por isso neutras e independentes dos resultados que delas se venham a extrair.

O método tradicional é claramente "source oriented" e o método Amorim claramente "result oriented". Não é correcto dizer em abstracto que um é melhor que outro. Ambos cumprem objectivos diferentes e destinam-se a cenários de utilização diferentes. Para operadores qualificados os resultados de cada um deles devem ser indistinguíveis e o método Amorim é mais rápido. Sempre que haja necessidade de controlar o processo mais de perto, ou haja razões para querer validar uma primeira reconstituição, então o método Fleury-Henry permite voltar às fontes sem voltar ao arquivo.

Esta dicotomia transpõe-se ampliada para o mundo da informática. Só que aqui, como a informação circula e se transforma mais rapidamente, as diferenças entre uma aproximação orientada para a fonte e uma aproximação orientada para os resultados é ainda mais significativa.

Outro objectivo a perseguir é a flexibilidade da representação de dados. Neste tipo de investigação encontramos necessariamente pessoas sobre as quais foi possível acumular uma grande quantidade de dados e outras sobre as quais se sabe muito pouco. É necessário que o modo como a informação é representada possa lidar com essa variação. Não podemos, neste tipo de pesquisa, operar com fichas pré-definidas de estrutura rígida[9].

---

[9] Sobre esta necessidade ver: Beveridge, Andrew; Sweeting, George- *Running Records and the Automated Reconstruction of Historical Narrative*. "Quantum/Historical Social Research", nº35, Jul 1985, p.31-44; Karweit, Nancy; Kertzer, D. - *Database management for life course family research,* in Current perspective in Aging and the Life Cycle, Jai Press Inc., 1986, vol.2 p.167-188.

O terceiro objectivo a exigir ao sistema é que suporte um processo de identificação reversível e tanto quanto possível automático[10]. A reversibilidade é um requisito de controle e verificação. Num estudo nominal a identificação das pessoas é talvez o aspecto de tratamento dos dados que mais determinante se torna nos resultados finais. Com efeito, erros de identificação podem ter consequências muito nocivas na interpretação dos resultados. Já os estudos que passam por uma fase de identificação de pessoas como passo intermédio para um estudo estatístico global não são tão sensíveis a erros de identificação. Por exemplo os programas usados pelo Grupo de Cambridge para reconstituição de famílias geram aleatoriamente uma identificação quando não existe informação suficiente para distinguir entre dois homónimos. Do ponto de vista da análise dos resultados finais esse procedimento é perfeitamente aceitável. Se duas pessoas têm o mesmo nome, são ambos solteiros, nasceram no mesmo dia, e moraram no mesmo lugar então, quando um deles morre, passado 30 anos, é indiferente do ponto de vista do demógrafo se foi um ou outro. A identificação pode ser feita ao acaso porque as variáveis pertinentes a extrair dos dados não são afectadas por escolhas desse tipo (duração média de vida, por exemplo, ou outro medidor demográfico).

Já o estudioso das pessoas em si, o nominalista, extrai muita informação do caso concreto, do trajecto individual. Uma identificação aleatória é aceitável em muito menos casos.

---

[10] Wrigley, E.A. (ed.), *Identifying People in the Past,* London. Edward Arnold, 1973 . Roger Schofield afirmava este requisito de modo enfático: *"If the judgements we make about specific links have any claim to intellectial respectability, we ought to be able to specify the principles on which they are based. If we can do that, we can express those principles in the form of a computer program and get the machine to implement them far more consistently than ourselves"* in Schofield, Roger, *Automatic Family Reconstitution*. "Historical Methods", volume 25, number 2, (Spring 1992), p. 75-79.

Sendo o processo de identificação determinante para os resultados deve por isso ser passível de ser revisto, feito e desfeito quantas vezes forem necessárias. Em nenhum caso devem os passos da identificação ficar definitivamente "invisíveis" no sistema. Perante qualquer biografia reconstituída ou qualquer resultado agregado o utilizador do sistema deve poder rever a informação original antes de ter sido consolidada pelo cruzamento nominal.

A automatização serve múltiplos propósitos: em primeiro lugar diminui a importância das escolhas fortuitas e ocasionais do investigador e reforça os procedimentos racionalizados e formalizáveis. É preferível que o modo de fazer do investigador seja incorporado em rotinas automáticas, cujos resultados podem ser revistos e afinados, a que esse saber se manifeste por decisões pontuais, "manuais", difíceis de reproduzir e de contextualizar mais tarde. A automatização permite a reproductibilidade de resultados num passo essencial do processo de investigação. Torna-se assim um mecanismo importante de validação. Para além disso, ao formalizar muito do que é o conhecimento intuitivo do historiador, torna transmissível e re-utilizável saberes que doutro modo dificilmente transbordariam para fora da prática do especialista.

Todos os que tiveram experiência deste tipo de pesquisa sabem que com o tempo se constrói na mente do investigador um conhecimento quase simbiótico do universo que estuda. As palavras roídas pelo tempo advinham-se a partir de algumas letras. Os nomes que os recém chegados consideram indecifráveis são claros para quem conhece de perto o universo nominal. Acontece que, ao ver apenas o nome da criança baptizada, se adivinha quem é o padrinho.

Este conhecimento que lentamente se consolida não é muito fácil de formalizar mas é aí que reside a base da perícia com que o investigador opera a reconstituição. É esse saber que o qualifica como perito. Os programas

informáticos que se comportam como "peritos" em determinado assunto designam-se por "sistemas periciais".

Os sistemas periciais procuram modelizar o conhecimento intuitivo resistente a formulações analíticas. Normalmente são construídos em interacção entre um engenheiro informático e um perito, e têm encontrado as suas aplicações mais célebres precisamente em áreas que envolvem uma grande quantidade de dados tratados por processos difíceis de racionalizar, onde a intuição e o hábito parecem jogar um papel determinante. É o caso da medicina e da prospecção geológica.

O último objectivo diz respeito às fases mais adiantados da investigação. Uma reconstituição de comunidade produz uma grande quantidade de informação cuja apreensão põe problemas não triviais. Para extrair sentido das biografias de centenas de pessoas, das genealogias intrincadas, das redes de relações, da lógica da ocupação do espaço, é necessário criar instrumentos de visualização específicos. Assim um sistema de RCH tem de possuir essa componente de visualização de resultados que terá de ser criativa e aproveitar bem as capacidades gráficas das máquinas de hoje.

Os próximos capítulos detalham as soluções encontradas para estes requisitos. De seguida abordaremos o problema da criação de uma estrutura de dados flexível que suporte informação variável no tempo e oriunda de fontes muitos diversas.

3.2. REPRESENTAÇÃO DA INFORMAÇÃO NOMINAL: FORMALISMOS, ESTRUTURAS E IMPLEMENTAÇÃO

3.2.1. INTRODUÇÃO

No capítulo anterior estabelecemos em linhas gerais as necessidades de um sistema informático para a Reconstituição de Comunidades Históricas. Neste capítulo abordaremos um primeiro elemento essencial desse sistema: o formalismo para a representação dos dados. Aqui explicitar-se-á a estrutura de uma base de dados que suporta informação sobre pessoas, e outro tipo de entidades históricas, recolhidas em várias fontes[11].

---

[11] Bernard, L - *Relational Theory, SQL and Historical Pratice* In Denley, Peter; Fogelvik, Stefan; Harvey, Charles (eds), *History and Computing II,* Manchester, Manchester University, 1989, p.63-71; Hartland, P.; Harvey, C. - *Information Engineering and Historical Database, ibidem,* p.44-62.

As características principais dessa estrutura são as seguintes:

- Recolha de informação dispersa por várias fontes mantendo a origem de cada informação;
- Suporte a um processo de identificação reversível: a informação sobre várias pessoas pode ser consolidada numa única biografia e o processo revertido se fôr necessário;
- Capacidade de lidar com um número variável de informações para cada pessoa (ou seja cada pessoa tem um número potencialmente ilimitado de informações a seu respeito);
- Representação da ligações entre pessoas ou entre qualquer outro tipo de entidade;
- Capacidade de tratar informações variáveis no tempo;
- Possibilidade real de implementação para grandes quantidades de dados;

Para entender o modo como se obteve um sistema que satisfaz estas características é necessário um enquadramento sobre o modo como, em informática, se constroem representações formais do mundo real.

3.2.2. FORMALISMOS PARA A REPRESENTAÇÃO DOS DADOS

3.2.2.1.Entidades, atributos, relacionamentos.

A variedade de informações que se podem desejar tratar informaticamente é infinita. Contudo é sempre possível reduzir o trabalho de encontrar um modelo para determinado conjunto de dados à definição do modo como se representam três aspectos distintos e fundamentais: entidades, atributos e

relações. Diferentes metodologias de programação ou diferentes modelos de representação de dados propõem soluções diferentes para estes três aspectos.

Para melhor compreender a explicação que se segue é necessário definir previamente o significado de determinados termos, que, sendo de uso corrente, assumem significados específicos no contexto da literatura sobre metodologias de programação e estruturas de dados.

Designamos por "entidade" algo que existe ou existiu e pode ser representado num computador. As entidades incluem objectos ou seres que realmente existiram (como uma determinada pessoa ou um determinado edifício), mas também construções abstractas como "família", "rede de relações", "acontecimento".

As entidades são descritas por "atributos". São atributos das entidades do tipo "pessoas" o nome, a data de nascimento, a profissão e a residência, entre outros. Alguns atributos assumem um único valor para determinada entidade concreta (data de nascimento, por exemplo), outros vão-se modificando ao longo do tempo (residência, profissão, idade, etc...). O que distingue duas instâncias de uma entidade uma da outra é o facto de terem valores diferentes para os mesmos atributos (diferentes nomes, diferentes idades, diferente sexo, etc...), embora alguns possam ter o mesmo valor.

O conjunto de atributos que distinguem uma entidade concreta de outras entidades do mesmo tipo denomina-se por "chave". A "chave" é o atributo, ou conjunto de atributos, que funciona como identificador dos indivíduos. Nos dias de hoje a "chave" natural da entidade "pessoa" seria um número de identificação oficial como o número do bilhete de identidade. As pessoas que viveram no passado não possuíam um atributo identificador tão claro. O nome raramente serve de atributo "chave" devido ao facto de várias pessoas terem o mesmo nome. Veremos mais tarde como se resolve esta dificuldade. O

importante é reter que para cada tipo de entidade deve haver um ou mais atributos que distinguem cada entidade concreta das outras.

Outra característica importante das entidades é que estas relacionam-se entre si. As pessoas estabelecem relações de vários tipos com outras pessoas (parentesco, solidariedade, inimizade...) e com outro tipos de entidades (possuem objectos, pertencem a instituições....). Tais como as entidades propriamente ditas, as relações entre entidades são qualificadas por atributos. Assim uma relação de parentesco entre duas pessoas é descrita por atributos como o nome do parentesco, o grau consaguinidade ou afinidade e a identificação das pessoas relacionadas.

Existem vários tipos de entidades numa investigação histórica deste tipo: pessoas, instituições, propriedades, etc... Cada tipo de entidade ocorre várias vezes: existem várias pessoas, várias instituições, várias propriedades.

Podemos por isso falar de pessoas em abstracto e falar de pessoas concretas. Ao dizermos, por exemplo, que todas as pessoas têm um atributo que é o sexo, que esse atributo pode assumir dois valores distintos e que não varia ao longo do tempo, estamos a falar abstractamente, ou genericamente, das entidades do tipo pessoa. Ao dizermos que a primeira testemunha da devassa de 1693 é uma mulher estamos a falar concretamente de uma pessoa específica. Existe assim uma dualidade entre o "nível abstracto" e o "nível concreto".

Essa dualidade é da mesma natureza que a diferença entre definir um formulário e preencher um formulário. A definição do formulário implica uma análise de quais os campos que devem estar presentes e quais os valores que podem receber. À definição do formulário está igualmente associada a ideia de "procedimento" O formulário é definido para determinado fim e desencadeará, uma vez preenchido, uma série de acções concretas que serão os "procedimentos" associados ao formulário. Embora essas acções ocorram

posteriormente ao preenchimento, terão sido, seguramente, definidas no mesmo processo de análise que definiu quais os campos a incluir no formulário.

Na análise de informação existe o mesmo duplo aspecto. Para cada entidade definem-se previamente quais os atributos possíveis e que tipo de valores pode receber. Definem-se igualmente quais as "acções" ou "procedimentos" que podem operar sobre essas entidades. Depois a informação é recolhida para esse "molde" e os procedimentos podem ser usados.

Os conceitos de "entidade", "atributo" e "relacionamento" são centrais no processo de análise do modo como vamos representar em computador a informação do mundo. Esse processo implica a identificação das "entidades, atributos e relacionamentos" num dado conjunto de informação. Esta operação chama-se "modelização da informação". Os conceitos que temos estado a referir pertencem a uma metodologia específica de modelização designada por "modelo entidade-relacionamento" (em inglês "Entity-Relationship Model"). Embora existam outras metodologias alternativas, cuja explicitação está fora do âmbito deste trabalho, o "modelo entidade relacionamento é o mais usado. As razões dessa popularidade são claras: é um modelo relativamente simples de entender e facilmente tradutível para as estruturas de bases de dados relacionais.[12]

O modelo entidade-relacionamento tem sido objecto de considerável investigação, quer teórica, quer aplicada e foi objecto de algumas modificações e extensões. A mais importante é conhecida por modelo "entidade-

[12] Lochovsky, Frederick (ed),*Entity-Relationship approach to database design and querying. Proceedings of the Eight Internacional Conference on Entity-Relationship Approach,* Amsterdam New York Oxford Tokyo: North-Holland, 1990.

relacionamento extendido" (em inglês: "Extended Entity Relationship Model" (EER). Do ponto de vista dos nossos objectivos o modelo extendido acrescenta um conceito muito importante que é o de especialização.

Por especialização designa-se a relação existente entre duas entidades quando uma é um caso particular de outra e corresponde à noção de que podemos classificar entidades em tipos e subtipos. Por exemplo: uma venda é um caso particular de uma escritura notarial. Um aforamento é outro caso particular de escritura notarial. Aforamento e venda têm um certo número de características comuns, que lhes vêm de ambas serem "escrituras notariais": foram elaboradas por um notário, num local determinado, numa data determinada. Por outro lado, tanto o aforamento como a venda têm atributos específicos do seu tipo de escritura: o valor da venda, as condições do foro.

Quando uma entidade especializa outra mantém os mesmos atributos que a entidade de nível mais geral e acrescenta alguns específicos do seu sub-tipo. No nosso exemplo a entidade "escritura notarial" teria como atributos o nome do notário, a cota do livro, a data e o local onde foi feita. Estes atributos são comuns a todas as escrituras notariais. Cada subtipo acrescenta então os seus atributos específicos ao conjunto de atributos comuns. As escrituras de venda terão atributos relativos ao objecto, ao preço e às condições de pagamento. Os empréstimos acrescentariam atributos relativos ao montante da dívida, juros, prazos de pagamento e garantias.

Devemos ainda frisar um facto importante. A relação entre a entidade "escritura notarial" e a entidade "escritura de venda", que chamámos de especialização, é uma relação de natureza diferente das relações que mais acima referimos que existem entre as entidades do mundo real: relações entre pessoas, entre pessoas e objectos, etc... Estas são relações "reais" que ocorrem no mundo concreto. A relação de especialização é uma relação abstracta, que estrutura a representação da informação mas que não corresponde a nenhum

facto real. Podem ocorrer relações reais e concretas entre escrituras. Por exemplo uma escritura notarial pode revogar outra anteriormente feita. Dizemos que existe uma relação de "revogação" entre as duas. Esta é uma relação real entre objectos ou acontecimentos reais. Mas quando dizemos que as entidades do tipo "escrituras de venda" têm uma relação de especialização com as entidades do tipo "escrituras notariais" estamos a construir uma relação abstracta entre entidades formais.

O conceito de especialização é muito importante no desenho de bases de dados flexíveis em situações de grande complexidade de informação, que é o caso da RCH. A especialização permite organizar as várias entidades numa hierarquia de tipos e sub-tipos. Esta organização do ponto de vista da estrutura dos dados permite uma construção simplificada dos programas que os vão trabalhar, como veremos mais adiante.

Recapitulando, os conceitos que utilizaremos para descrever a informação envolvida na reconstituição de comunidades históricas são relativamente poucos: entidades, relações, atributos e especializações. Estes conceitos aplicados aos dados da comunidade fornecerão uma visão estruturada e hierárquica dos vários tipos de informação. Só uma visão desse tipo permite a passagem à fase de implementação.

3.2.2.2.A base relacional

Uma estruturação puramente lógica da exposição levar-nos-ia agora à aplicação destes conceitos à informação de uma comunidade histórica e à definição da hierarquia formal de representação. De facto os conceitos fundamentais do ponto de vista formal estão explicitados. Sabemos contudo que a formalização é muito influenciada pelas condições concretas de implementação. Isto significa que não se constrói um sistema formal de

representação sem ter ideias concretas de como se vai implementar esse sistema. Seria prematuro, por isso, começar neste momento a identificar entidades, relações e atributos. Se assim fosse, muitos aspectos da representação que foi criada por este trabalho pareceriam arbitrários ou injustificados. Uma das características dos métodos aqui descritos é que eles procuraram ser formalmente minimalistas, só detalhando as representações quando a prática concreta e as exigências técnicas o impunham. É necessário por isso adiantar ainda mais algumas considerações, agora ao nível do modo como os modelos do tipo "entidade - relacionamento" são implementados por programas de computador. Com essa informação adicional podemos então passar à análise concreta de um sistema de informação para a reconstituição de comunidades históricas.

Ao longo do desenvolvimento das técnicas informáticas foram sendo criadas várias soluções que procuraram preconizar formas de lidar com os conceitos do modelo entidade-relacionamento e dar-lhes uma implementação técnica fiável. Foram propostos métodos de análise da informação real e modos como os resultados da análise podem ser transportados de forma clara e inequívoca para os computadores. Essas técnicas estão intimamente ligadas ao desenvolvimento de linguagens de programação e de sistemas de gestão de bases de dados[13].

Os sistemas de bases de dados são programas informáticos que foram concebidos para gerir grandes quantidades de informação. Normalmente distinguem-se três passos essenciais na implementação de um sistema de gestão de base de dados (SGBD):

---

[13] Como introdução geral a este tema: Elmasri, Ramez; Navathe, Shamkant B., *Fundamentals of Database Systems*. RedWood Cliffs: Benjamin Cummings Publishing Company, s/d e Korth, Henry F.; Silberschatz, Abraham, *Database System Concepts*. New York, MacGraw Hill, 1991.

- A definição da estrutura dos dados que se pretendem tratar.
- A introdução e modificação de dados na estrutura definida.
- A consulta da informação e obtenção de resultados.

Os SGDB classificam-se conforme o modelo que usam para representar a informação. É comum distinguir os modelos relacional, hieráquico, de rede e, mais recentemente, o modelo orientado a objectos. Estes modelos dizem respeito à *implementação*, o modo como a informação é definida, introduzida e consultada. São por isso independentes de um modelo de *análise* como o modelo entidade-relação, que enquadra o processo de compreensão da estrutura dos dados. Um dado caso real pode ser descrito segundo os conceitos do modelo entidade-relação e implementado alternativamente em vários sistemas de bases de dados que sigam modelos diferentes de implementação. Em princípio são as características dos dados obtidas pela aplicação do modelo de *análise* que determinam qual o modelo de *implementação* a usar.

Na prática o modelo relacional de implementação é quase sempre usado em projectos de grande dimensão onde os dados têm uma estrutura regular. Assim se passa no nosso caso. As razões de escolha de sistemas baseados no modelo relacional têm a ver com as características formais bem conhecidas desse modelo, nomeadamente o modo como transcreve uma análise do tipo "entidade-relação", assim como a disponibilidade de programas fiáveis. Estes aspectos serão explicitados a seguir.

O modelo relacional assenta sobre três atributos essenciais:

- Uma metodologia de análise da informação que produz uma representação da mesma passível de ser tratada por programas informáticos.

Essa metodologia é baseada em formalismos que garantem a acessibilidade da informação e a independência em relação à implementação concreta.

- Define uma linguagem universal para a definição da estrutura da informação, a sua manipulação (inserção e eliminação de informação) e a pesquisa de dados concretos. Outros modelos de base de dados, por não terem claramente definidos estes aspectos formais e linguísticos, nem sempre garantem um acesso a todos os items de informação.

- Os programas necessários para implementar concretamente o modelo relacional existem para praticamente todo o tipo de computadores. Esses programas chamam-se "sistemas de gestão de base de dados relacionais" (SGBDR) e têm larga utilização no mundo empresarial e académico, cumprindo tarefas de gestão de enormes quantidades de informação. Em consequência da sua utilização generalizada, os sistemas de gestão de base de dados relacionais com maior quota no mercado são programas robustos e de grande fiabilidade.

O modelo relacional representa a informação sob a forma de tabelas[14]. Uma tabela é a forma sob a qual se verte a informação sobre uma entidade. Imaginemos uma entidade do tipo pessoa, para a qual se registam os atributos: nome, data de nascimento e data de morte. Esta entidade seria representada por uma tabela denominada "pessoas" com três colunas: "nome", "data de nascimento", "data de morte". Cada linha da tabela representa uma pessoa concreta. Cada coluna representa um atributo. O conceito de "chave", que vimos anteriormente, é fundamental no modelo relacional. Cada tabela deve ter

[14] A explicação que se segue tem como objectivo apresentar o essencial do modelo relacional de modo que as opções que foram tomadas em termos de desenho da base de dados possam ser entendidas. A terminologia utilizada e a simplificação de determinados aspectos são fruto de uma escolha consciente de tornar estes conceitos acessíveis aos não especialistas.

uma ou mais colunas que em conjunto assumem um valor diferente em cada linha da tabela e que funcionam como identificador.

As relações entre entidades são representadas no esquema relacional diferentemente conforme as características das relações. Se a relação fôr de um para um para um ou de um para muitos, adiciona-se numa das tabelas os atributos que identificam as entidades da tabela relacionada Se se trata de relações muitos para muitos cria-se uma nova tabela para representar a relação. Essas tabelas possuem colunas que identificam as entidades concretas que se estão a relacionar e colunas adicionais para os atributos da relação. Por exemplo uma tabela que registe as relações de parentesco entre pessoas deverá ter as colunas necessárias para identificar as duas pessoas em causa e uma coluna onde se regista o tipo parentesco que as une.

Uma das regras importantes do modelo relacional é a do caracter atómico dos atributos. Cada "célula" da tabela, que representa um atributo de uma entidade concreta, deverá assumir um único valor, e esse valor não deve ser composto. Assim, por exemplo, se tivermos uma tabela representando pessoas com uma coluna designada "profissão", não deveremos introduzir mais do que uma profissão em cada linha da tabela. Existem outros processos, como veremos, para representar correctamente atributos que assumem vários valores ou que possuem outra informação associada. Introduzir valores compostos, ou variações de um valor, numa mesma célula de uma tabela, é um erro infelizmente frequente no desenho de ficheiros para investigações históricas.

Resumindo as regras do modelo relacional:

- cada entidade uma tabela.
- cada atributo uma coluna.

- cada linha da tabela representa uma ocorrência concreta da entidade que a tabela representa.
- cada atributo assume um valor atómico.
- cada relação entre entidades é representada por uma tabela que inclui as chaves das entidades relacionadas e os atributos da relação[15].

Estas regras aplicadas a uma análise em termos de entidades-relações permitem-nos propôr uma estrutura de base de dados capaz de realizar os objectivos enumerados anteriormente. Veremos de seguida a estrutura da base de dados relacional criada.

3.2.3. A ANÁLISE E IMPLEMENTAÇÃO DA INFORMAÇÃO NOMINAL.

3.2.3.1.A informação biográfica

Comecemos pelo núcleo duro de informação RCH: a informação associada às biografias individuais[16]. Os nossos requisitos neste domínio incluíam:

- Registo flexível de informação variável com o tempo.
- Registo da origem (fonte ou documento) de cada informação.

---

[15] Na verdade, certo tipo de relações dispensam a criação de tabelas próprias, como ficou explicado acima.

[16] Abordagens semelhantes, conceptualmente, á que aqui seguimos, especialmente no que toca à separação entre os atributos variáveis e a informação estável das pessoas encontram-se em: Beveridge, Andrew; Sweeting, George- *Running Records and the Automated Reconstruction of Historical Narrative. cit;* Karweit, Nancy; Kertzer, D. - *Database management for life course family research, cit* .

- Suporte para um processo reversível de identificação.

O primeiro passo do processo de análise consiste em identificar as entidades, atributos e relações envolvidas na informação biográfica sobre as pessoas. A entidade central deste processo é a entidade "pessoa". A entidade pessoa possui um conjunto muito variado de atributos: nome, sexo, profissão, idade, residência, etc... Por outro lado as pessoas relacionam-se entre si de muitas variadas maneiras: relações de parentesco, vizinhança, solidariedade, e muitas outras.

Primeira questão: quais os atributos que podem funcionar como "chave", ou identificador, da entidade "pessoa"?. Como estamos a tratar de indivíduos anteriores à criação pelo Estado de mecanismos de identificação unívocos, como o bilhete de identidade, temos dificuldade em encontrar um atributo, ou mesmo um grupo de atributos, que identifique sempre uma pessoa concreta. De facto sabemos que o nome, embora possa identificar sem ambiguidade algumas pessoas de uma população, não serve de atributo distintivo para a maior parte dos indivíduos, devido aos elevados índices de homonomia existentes. Não podemos tão pouco, num sistema genérico, usar combinações como nome-residência-data de nascimento, ou outras semelhantes, porque a diversidade das fontes nem sempre garante a presença dos vários atributos. Contudo temos de encontrar uma forma de referir univocamente as pessoas que encontramos. Trata-se de uma exigência formal, como vimos, dos modelos de análise e de implementação, que se justifica muito intuitivamente: se num determinado momento do processo de identificação necessito de designar que determinada pessoa é a mesma que outra, como "nomeio" essas pessoas sem ambiguidades, em populações em que mais de 10% dos indivíduos têm exactamente o mesmo nome?

A solução encontrada, que aliás é comum em casos semelhantes, é de criar artificialmente um atributo identificador no momento em que as referências às pessoas são registadas. Esse atributo identificador assume a forma de uma espécie de matrícula alfa-numérica. Na nossa implementação concreta do processo de registo de dados essa matrícula é gerada automaticamente e é mantida invisível do utilizador até ao momento em que é necessário referir explicitamente alguém. Chamamos a esse atributo artificialmente criado o "id" da pessoa (abreviatura de "identificação").

Uma segunda questão relacionada com os atributos das pessoas diz respeito à forma como os atributos variam diferentemente com o tempo. Certos atributos são estáveis e definitivos, como por exemplo a data de nascimento, a data de morte, o sexo. Outros variam com o tempo e cada valor que assumem só pode ser entendido em função de um determinado momento histórico: idade, profissão, estado civil, residência, e praticamente todos os outros atributos que as fontes vão revelando sobre os indíviduos. Para os atributos que variam no tempo interessa-nos registar a data em que determinado valor se verificou assim como a fonte que forneceu a informação. Isto significa que a nossa informação sobre os atributos é relativamente complexa. Na verdade temos atributos sobre os atributos (data em que o valor se verificou, fonte, etc...). Dentro do quadro formal que nos guia aqui esta complexidade inerente a certo tipo de atributos significa que temos que tratar esses atributos como se fossem entidades, dependentes, é certo, da entidade principal "pessoa" mas com um conjunto de valores próprios. Iremos ter por isso um segundo tipo de entidade, ligada à entidade "pessoa" a que chamaremos "atributo". Na literatura do modelo entidade-relação designa-se este segundo tipo de entidades por entidades "fracas" uma vez que a sua função é registar informação que descreve uma entidade "forte".

Temos assim duas entidades fortemente interligadas: a entidade "pessoa" com atributos fixos como: "id", "sexo", "data-nascimento", "data-morte"; a entidade "atributo" regista os atributos variáveis no tempo e recolhidos em fontes diversas. A entidade "atributo" tem atributos como: "id" da pessoa a que o atributo se refere, o tipo de atributo que se trata (residência, idade, profissão...), o valor que o atributo assume num momento concreto (um lugar ou rua, uma idade, uma profissão...), a data correspondente a esse momento e a fonte onde a informação apareceu.

A implementação concreta, sob os auspícios do modelo relacional, afasta-se relativamente pouco desta análise. Restringimos os atributos fixos da entidade "pessoas" a "id" e "sexo", registando as datas de nascimento e morte como se fossem atributos variáveis (é útil poder registar a fonte em que essa informação originou) e acrescentamos o atributo "nome", que obviamente não é um atributo fixo, mas que é útil ficar registado na entidade "pessoa" pelo valor mnemónico que tem. Acrescentamos ainda, como será regra em todo o tipo de entidades do sistema, um campo genérico de observações. Quanto à entidade "atributos" definimos as seguintes colunas: o "id" da pessoa, nome do atributo, valor do atributo, data em que valor se registou e identificação da fonte. Temos ainda que resolver o problema seguinte: as entidades "atributos", como todas as entidades, têm de ter uma "chave" que neste caso tem de identificar univocamente um registo de atributo dentro de todos os registos de atributos da base. Tal como fizemos para as pessoas vamos utilizar um identificador artificial, uma matrícula gerada automaticamente no momento do registo e tratar de a esconder do utilizador sempre que possível. Temos em consequência, e até agora, duas tabelas (ver quadros 3.1 e 3.2).

Quadro 3.1: Tabela para registo de pessoas

Pessoas

Atributos/colunas	*Explicação*
id	matrícula identificadora (automática)
nome	nome
obs	observações gerais

Quadro 3.2 Tabela para registo de atributos de pessoas

Atributos

Atributos/colunas	*Explicação*
id	matrícula identificadora (automática)
pessoa	id da pessoa respectiva (refere a coluna "id" da tabela "pessoas")
atributo	tipo de atributo
valor	valor do atributo
data	data do atributo
fonte	fonte original (código)
obs	observações gerais

A observação das duas tabelas 3.1 e 3.2 permite-nos introduzir as seguintes regras: todas as tabelas no sistema terão sempre um atributo que funciona como chave e que será sempre uma matrícula gerada automaticamente (indicamos esse atributo/coluna sublinhando o respectivo nome). Em raríssimos casos recorreremos a um atributo real para obter a chave de uma entidade. A criação de chaves artificiais poupa o trabalho de procurar combinações complexas de atributos que nem sempre seriam funcionais. Aliás, de modo geral, os sistemas de gestão de bases de dados relacionais não

facilitam a operação sobre tabelas que incluem chaves de identificação criadas por agregação de atributos. Todas as operações que operam sobre essas chaves são mais difíceis de expressar na linguagem padrão desses sistemas. Decidimos também designar sempre o atributo chave por "id". Finalmente o processo de geração automática de matrículas de identificação que utilizamos garante que todas as matrículas são diferentes independentemente da tabela em que ocorrem. Note-se que esta característica é uma extensão nossa ao requisito formal da univocidade da chave. O modelo relacional apenas exige que as chaves identifiquem univocamente os indivíduos de determinada entidade, ou seja, que o atributo que serve de chave seja diferente para cada linha de uma tabela. A exigência de univocidade geral das chaves, ou seja que o atributo "id" de cada entidade nunca assume um valor repetido, independentemente da entidade, permitirá, como veremos mais adiante, efectuar algumas operações de generalização e especialização interessantes.

Vejamos um exemplo concreto para seguidamente passarmos a uma análise mais detalhada das características destas tabelas.

Suponhamos o seguinte fragmento de um documento:

```
[Na devassa de Soure de 1692, em 4 de Outubro, foi interrogada a
testemunha]

 José Machado, solteiro, da vila, que de idade disse ter 24 anos....
```

No nosso esquema este fragmento dá origem, nas tabelas de pessoas e de atributos, às linhas que se podem ver nas quadros 3.3 e 3.4.

Quadro 3.3: Linhas na tabela de pessoas (I)

Pessoas

id	*nome*	*sexo*	*obs*
d1692-t1	jose machado	m	

Quadro 3.4 Linhas na tabela de atributos (I)

Atributos

id	*pessoa*	*atributo*	*valor*	*data*	*fonte*	*obs*
d1692-t1a1	d1692-t1	ec	s	16921004	ldp1692	
d1692-t1a2	d1692-t1	residencia	soure	16921004	ldp1692	"vila"
d1692-t1a3	d1692-t1	idade	24	16921004	ldp1692	

As colunas "id" de cada uma das tabelas contêm as matrículas identificadoras. Como se referiu acima, na utilização corrente do sistema estas matrículas são mantidas relativamente invisíveis. No caso das matrículas de identificação dos atributos ("d1692-t1a1", "d1692-t1a2",...) a sua existência cumpre uma função puramente interna e não chegam a ser mostradas aos utilizadores. Quanto às matrículas que identificam ocorrências de pessoas a sua visibilidade é maior porque o utilizador frequentemente terá que designar uma pessoa concreta de forma não ambígua através da sua matrícula. As matrículas podem ter qualquer sequência de caracteres embora o sistema procure gerar designações com algum sentido mnemónico. Neste exemplo a matrícula "d1692-t1" recorda que se trata da testemunha um da devassa de 1692. O número máximo de caracteres de uma matrícula é de 42.

As informações são registadas sempre em minúsculas para evitar os problemas que por vezes surgem com variações fortuitas. Uniformizar o registo sempre em maiúsculas ou sempre em minúsculas acelera a introdução de dados, simplifica a consulta e reduz a possibilidade de erro. As letras minúsculas são mais legíveis que as maiúsculas e por isso serviram de base de uniformização. Esta restrição do uso de minúsculas não se aplica aos campos

de observações das tabelas, onde, normalmente, é entrado texto para posterior leitura e não para processamento automático.

Os nomes foram registados na sua forma documental, passando apenas por uma actualização ortográfica que se aplicou a todos os items de informação recolhidos. Mantivemos nos nomes as partículas de ligação ("de", "e", "o", "a"...) porque, sendo só frequentes em nomes complexos, normalmente apanágio das classes mais elevadas, poderiam encerrar particularidades de nomeação pertinentes no processo de identificação. Tal não foi o caso e acreditamos actualmente que as partículas de ligação dos nomes podem ser ignoradas sem perda de informação.

Colunas como "sexo", na tabela de pessoas, e "atributo" na tabela de atributos só admitem termos escolhidos de listas relativamente curtas. São colunas de vocabulário controlado "a priori", ou seja, é possível determinar em cada momento da investigação quais os termos que podem ser introduzidos nestes campos. O leque de vocábulos da coluna "atributo" é curto, consta apenas de 17 termos entre os quais: "ec" (estado civil), "residência", "naturalidade", "idade", "profissão", "cargo", "titulo" [17]. O sexo é representado por dois termos ("m" e "f").

Colunas como "valor" na tabela de atributos têm igualmente um vocabulário finito mas de controle mais difícil. É comum, sobretudo com o atributos como "residência" ou "naturalidade", que se registem formas diferentes do que mais tarde se descobre ser o mesmo topónimo. Enquanto que com o "nome" dos atributos o vocabulário é pequeno e alterável por decisões conscientes e pontuais do investigador, já o "valor" desses atributos só é uniformizável à posteriori. No caso de Soure, o leque de valores definitivos para os topónimos só estabilizou definitivamente durante o processo de identificação das pessoas,

[17] Ver apêndice para a lista completa e totais de ocorrências.

pois só nessa altura foi possível resolver dúvidas referentes a formas diferentes para o mesmo topónimo, retraçando as pessoas a elas associadas[18].

As datas foram registadas sob a forma de valores de 8 dígitos, sendo os primeiros quatro correspondentes ao ano, os dois seguintes ao mês e os dois últimos ao dia. Optámos por não utilizar os mecanismos próprios dos sistemas de gestão de base de dados para tratar com datas. A experiência demonstra que os programas informáticos de origem comercial implementam os mecanismos de calendário de forma muito simplificada, de modo que é preferível lidar com datas como se fossem números. É comum, por exemplo, que programas comerciais ignorem as regras sobre anos bissextos e anos de século e que limitem as datas que podem tratar às do século XX. No nosso caso, como recorremos a programas comerciais para a gestão da base de dados foi decidido utilizar o formato numérico descrito. A sequência de 8 dígitos que tem a vantagem de manter a ordem correcta durante os procedimentos de ordenação. Usámos ainda a convenção de introduzir zeros quando determinada parte da data (ano, mês ou dia) não era conhecida.

A coluna "fonte" da tabela de atributos indica de forma codificada a origem da informação. A cada fonte foi atribuída igualmente uma matrícula identificadora. Note-se que neste sistema não registamos a posição da informação dentro da fonte (a página, ou fólio). Essa omissão tem duas justificações. Por um lado aumentava grandemente a morosidade do registo de dados, acrescida da dificuldade suplementar de muitas fontes não possuírem

[18] Assim, em Soure, na passagem do século XVII para o XVIII os termos "Casal do Grisoma", "Casal da Ribeira", "Casal de Miguel Gante" e "Casal do Gante" designam todos o mesmo topónimo. A identificação do topónimo foi, neste caso, uma consequência do processo de identificação das pessoas que nele viviam, que por sua vez, faz parte duma das tarefas logicamente mais tardias do processo de RCH. Assim o controle *a priori* do vocabulário dos topónimos é impossível, sendo necessário adiar até à fase em que toda a informação está introduzida para proceder às identificações e uniformizações possíveis.

páginas numeradas originalmente. Por outro lado, o método de registo fragmenta a fonte em documentos ou actos e é relativamente rápido localizar com precisão a informação original. No nosso exemplo, apesar de não estar registado o fólio em que se obteve a informação sobre os atributos de José Machado, será muito fácil localizar na fonte a primeira testemunha da devassa de 1692. Para outros tipos de documentos utilizamos um processo semelhante: localizamos os documentos dentro de um livro e utilizámos a própria estrutura interna da fonte para rapidamente encontrar a informação original, quando é necessário. A necessidade de registar a posição de cada item de informação dentro da fonte só se põe para originais bastante volumosos que não possuam uma estrutura interna clara, quer cronológica, quer de outro tipo, o que nunca aconteceu na documentação que tratámos.

Finalmente uma referência às colunas de observações ("obs"). Normalmente essas colunas estão reservadas para a introdução de texto livre com notas que servirão apenas para futura referência do investigador. Trata-se de informação que, pela sua forma livre e não estruturada, não é sujeita a nenhum tratamento especial por parte dos programas que processam a informação. No exemplo acima serviu a coluna "obs" para registar a forma original do atributo residência que, no documento, se lia "vila" e que foi registada como "soure". De facto o valor a registar é "soure" porque é esse o significado de "vila" no contexto do documento particular que se está a processar. Interessa, por vezes, registar a forma original e a coluna "obs" pode ser utilizada para isso. Como veremos mais tarde este tipo de observações é introduzida automaticamente pelos programas que processam o registo de dados.

Às duas tabelas apresentadas até agora, "pessoas" e "atributos", iremos agora juntar uma terceira que registará as relações entre pessoas.

Retomando o exemplo anterior, da devassa de 1692, introduzamos um elemento adicional que consiste no facto de sabermos que José Machado é filho de Manuel Fernandes, de alcunha "o ratinho".

`José Machado, solteiro, da vila, que de idade disse ter 24 anos, *filho de Manuel Fernandes o ratinho....*`

Registar a informação adicional implica:

- Registar a nova pessoa, Manuel Fernandes, na tabela das pessoas
- Registar o seu atributo "alcunha" e o respectivo valor na tabela de "atributos"
- Registar a relação de parentesco pai/filho entre José Machado e Manuel Fernandes.

As tabelas anteriores (quadros 3.3 e 3.4) recebem assim uma linha adicional cada uma (ver quadros 3.5 e 3.6).

Quadro 3.5: Linhas na tabela de pessoas (II)

Pessoas

id	*nome*	*sexo*	*obs*
d1692-t1	jose machado	m	
d1692-r1	*manuel fernandes*	*m*	

Quadro 3.6: Linhas na tabela de atributos (II)

Atributos

id	*pessoa*	*atributo*	*valor*	*data*	*fonte*	*obs*
d1692-t1a1	d1692-t1	ec	s	16921004	ldp1692	
d1692-t1a2	d1692-t1	residencia	soure	16921004	ldp1692	"vila"
d1692-t1a3	d1692-t1	idade	24	16921004	ldp1692	

A relação de parentesco exige uma nova tabela, que servirá para registar todas as relações, de qualquer tipo, entre todas as pessoas. A tabela das relações inclui colunas que designam as duas pessoas envolvidas, o tipo de relação entre elas, a data em que a relação foi detectada e a fonte que forneceu a informação (quadro 3.7).

Quadro 3.7: Tabela para registo de relações

Relações

Atributos/colunas	*Explicação*
id	matrícula identificadora (automática)
origem	id da pessoa sujeito (origem) da relação
destino	id da pessoa objecto (destino) da relação
tipo	tipo da relação
valor	valor da relação
data	data da relação
fonte	fonte original (código)
obs	observações gerais

Quadro 3.8: Linhas na tabela de relações (relação pai/filho)

Relações

id	*origem*	*destino*	*tipo*	*valor*	*data*	*fonte*	*obs*
d1692-r1r1	d1692-r1	d1692-t1	parentesco	pai	16921004	ldp1692	

A coluna "id" desta tabela recebe uma matrícula, sempre diferente, automaticamente gerada. As colunas "origem" e "destino" encerram as

matrículas de identificação das duas pessoas envolvidas na relação. Note-se que uma relação do tipo "paternidade" é uma relação reflexiva: se X é pai de Y então Y é filho de X. Neste sistema não é necessário registar as duas formas. Os programas foram elaborados para compreenderem que determinadas relações são reflexivas. É por isso indiferente registar uma relação de paternidade como "X é pai de Y" ou "Y é filho de X". Neste exemplo registámos a relação "pai" e não "filho" pelo que temos de preencher as colunas origem e destino em conformidade: a origem é a matrícula do pai e o destino a matrícula do filho. Se tivessemos antes registado a relação como "filho" e não "pai" teríamos que inverter os valores da colunas "origem" e "destino" (quadro 3.9).

Quadro 3.9 Linhas na tabela de relações (relação filho/pai)

Relações

id	*origem*	*destino*	*tipo*	*valor*	*data*	*fonte*	*obs*
d1692-r1r1	**d1692-t1**	**d1692-r1**	parentesco	**filho**	16921004	ldp1692	

As relações têm um "tipo" e um "valor". O "tipo" distingue níveis de relações ("parentesco", "sociabilidade", "económicas"). O "valor" especifica a relação concreta. No caso do parentesco o vocabulário da coluna "valor" é dado pelo vocabulário normal utilizado para designar parentes (admitindo, contudo, formas para designar o grau de consaguinidade ou afinidade quando não conhecemos o parentesco concreto). Para outro tipo de relações a coluna "valor" utiliza um vocabulário controlado a priori.

As três tabelas pessoas/atributos/relações constituem o essencial do formalismo de representação da informação biográfica. Este esquema decompõe a entidade abstracta "pessoa" em três entidades operacionais inter-

relacionadas. Apesar da simplicidade formal da solução obtemos uma estrutura que permite o registo de uma quantidade variável de informação sobre cada indíviduo, e uma quantidade igualmente variável de relações entre indíviduos. Permite igualmente, do ponto de vista da implementação, obter rapidamente respostas a interrogações frequentes como sejam: quais os nomes das pessoas com determinado atributo, quais as relações registadas de uma pessoa. Não há limite para o número de atributos e de relações associadas a uma pessoa (acrescentam-se linhas nas respectivas tabelas conforme fôr necessário).

É importante reter que a nossa tabela de pessoas regista "ocorrências" de pessoas e não pessoas "reais". Quer isto dizer que cada referência a uma pessoa numa fonte gera as linhas correspondentes nas tabelas pessoas-atributos-relações. Frequentemente dar-se-á o caso de outras referências a essa pessoa se encontrarem já nessas mesmas tabelas. Assim o número de linhas existente na tabela "pessoas" é sempre muito superior ao número de pessoas reais na população que se estuda. O objectivo do processo de identificação é precisamente agrupar as linhas da tabela de pessoas em função das pessoas reais. Chama-se a esse processo "ligar os registos" ou, em inglês, "record linking".

O processo de identificação será objecto de uma explicação detalhada em capítulo posterior e aí serão detalhadas as estruturas de dados que registarão este tipo de informação. Neste momento importa continuar a expandir o esquema de registo da informação nominal passando àquilo que designámos por informação funcional.

3.2.3.2.A informação funcional: pessoas e actos

Retomemos o nosso exemplo baseado na devassa de 1692:

> José Machado, solteiro, da vila, que de idade disse ter 24 anos, *filho de Manuel Fernandes o ratinho....*

Vimos acima como a informação nominal associada às duas pessoas é registada num conjunto de três tabelas (pessoas, atributos e relações). Existe contudo um resíduo de informação não registado que diz respeito à *função* que ambas as pessoas assumem na fonte. José Machado é uma testemunha de devassa enquanto que Manuel Fernandes tem uma função acessória de auxiliar a identificação da testemunha. Esta informação tem de ser registada de algum modo.

Para registarmos o aspecto funcional da ocorrência de uma pessoa numa fonte iremos introduzir uma entidade adicional que denominaremos por "acto". Um "acto" é uma ocorrência, situada no espaço e no tempo, que deu lugar a um registo documentado envolvendo várias pessoas. Concretamente, nesta investigação, consideramos "actos" as devassas das visitas pastorais, os baptismos, casamentos, óbitos e escrituras notariais, que constituem o núcleo inicial da documentação tratada informaticamente. A estes podemos acrescentar no mesmo espírito medições e reconhecimentos em tombos, actas camarárias, processos inquisitoriais e praticamente qualquer tipo de acontecimento que gere um documento escrito. As escrituras notariais encerram, só por si, uma grande diversidade de actos.

A cada acto estão associadas várias pessoas com funções diferentes. No nosso exemplo, José Machado assume a função de "testemunha" num acto do tipo "devassa". Na mesma devassa outras pessoas assumirão a função de testemunhas, outras de acusados e outras ainda, como o pai de José Machado, a função periférica de serem "referidos" no acto sem que nele participem directamente.

Apesar da sua diversidade os actos têm um número de características comuns que nos servirão de base para a formalização. Nomeadamente todos os actos ocorrem numa determinada data. Actos de tipo diferente têm atributos diferentes mas iremos por agora contentarmo-nos em registar as características básicas dos actos, o que se fará por uma tabela com as características do quadro 3.10. O registo da devassa do nosso exemplo introduzirá uma linha nessa tabela, conforme se vê no quadro 3.11.

Quadro 3.10: Tabela para registo de actos

Actos

Atributos/colunas	*Explicação*
id	matrícula identificadora (automática)
data	data do acto
tipo de acto	devassa, baptismo,...
obs	observações gerais

Quadro 3.11: Linhas na tabela de actos

Actos

id	*data*	*tipo de acto*	*obs*
dev1692	16921004	devassa	

Existem várias opções para, a partir desta base, registarmos a informação funcional ligada às pessoas. Uma dessas opções consiste em criar uma nova tabela, semelhante à tabela de relações, que funcione como ligação entre as pessoas e os actos, registando a função que cada uma tem. Versões iniciais

deste sistema utilizavam um esquema desse tipo[19]. Contudo, com o evoluir da utilização decidimos simplificar o formalismo a partir de uma generalização do conceito de "relações". Vimos anteriormente como se registavam relações entre pessoas (quadros 3.7, 3.8 e 3.9). Alargaremos agora o conceito de "relações "para incluirmos na tabela respectiva não só relações inter-pessoais mas também relações entre pessoas e outro tipo de entidades, nomeadamente "actos". Em nada necessitamos de alterar a tabela representada pelo quadro 3.7. Basta-nos admitir o registo de relações do tipo "função em acto", cuja origem é uma pessoa e cujo destino é um acto.

Deste modo o registo da informação funcional do nosso exemplo consiste em acrescentar duas linhas novas à tabela do quadro 3.7 (ver agora o quadro 3.12): uma referindo que José Machado está relacionado funcionalmente como "testemunha" com a devassa de 1692 e que Manuel Fernandes está relacionado com o mesmo acto com a função de "referido" (termo que utilizamos para designar a função de pessoas que não participam directamente no acto mas são nele referidas, normalmente como elemento de identificação auxiliar dos participantes directos).

Quadro 3.12 Linhas na tabela de relações (incluindo relações funcionais)

Relações

id	*origem*	*destino*	*tipo*	*valor*	*data*	*fonte*	*obs*
d1692-r1r1	d1692-r1	d1692-t1	parentesco	pai	16921004	ldp1692	
d1692-t1r1	*d1692-t1*	*dev1692*	*funcao em acto*	*teste-munha*	*16921004*	*ldp1692*	
d1692-r1r2	*d1692-r1*	*dev1692*	*funcao em acto*	*referido*	*16921004*	*ldp1692*	

[19] Nomeadamente a estrutura de dados descrita em Carvalho, Joaquim Ramos de - *Soluzioni informatiche per microstorici*. "Quaderni storici". Vol. 78 (1991), p. 761-792.

A tabela de "relações" generalizou-se por este processo. Deixou de ser uma forma de registar relações entre pessoas para passar a registar relações entre entidades de natureza diversa, neste exemplo pessoas e actos. Esta generalização simplifica bastante o formalismo de base e podemos levá-la ainda mais longe assumindo que nesta tabela podem ser introduzidas relações entre qualquer tipo de entidades. Poderíamos imaginar o registo de uma relação entre duas escrituras, em que uma invalida a outra. Na origem da relação estaria um acto, no destino outro acto, a relação seria do tipo "entre actos" e o valor "derrogação". Significa isto que os valores das colunas "origem" e "destino" referem-se a entidades descritas em tabelas diversas ("pessoas" e "actos" no exemplo dado).

Esta generalização sugere outra. Pessoas e actos, embora no mundo real sejam entidades muito diferentes, assumem dentro do formalismo que estamos a apurar, alguns pontos de identidade. Acabámos de ver que ambas podem ser origem e destino de relações. Poderíamos também descrever os actos por atributos adicionais que seriam inseridos como linhas da tabela de atributos (quadro 3.2). Mais uma vez não necessitaríamos de modificar em nada a estrutura de dados já definida mas apenas generalizar a aplicação das tabelas.

Embora possa não ser evidente para o leitor, estes processos de generalização formal simplificam em muito a elaboração de programas e todas as tarefas relacionadas com a manutenção das bases de dados, incluindo, em última análise, a sua utilização pelo investigador. De um modo geral formalismos regulares e simples permitem a criação de programas que comunicam com os utilizadores de forma regular e simples, de modo que as nossas preocupações aqui não são puramente tecnicistas. Permitem, com um conjunto de tabelas relativamente reduzido, registar informações muito variadas sobre objectos históricos de tipos muito diferentes. A principal

vantagem que se extrai da utilização da generalização é a de tornar a estrutura da base de dados aberta à inclusão de novos tipos de dados com modificações mínimas. Em qualquer sistema de base de dados é sempre mais fácil introduzir novos dados em estruturas existentes do que adicionar novas estruturas sob a forma de mais tabelas. Como vimos acima, foi possível adicionar a informação funcional ligada às pessoas mantendo as tabelas existentes e deixamos em aberto a possibilidade de registar outro tipo de situações mais complexas, como a derrogação de um acto por outro.

A conclusão a que chegamos neste ponto é que a estrutura de três tabelas (pessoas-atributos-relações) que definimos para registar a informação biográfica pode ser generalizada para registar informação sobre tipos de entidades muito variadas. De seguida iremos explorar este aspecto um pouco mais profundamente e vermos como se pode, com um número de tabelas reduzidas, criar uma infra-estrutura geral e flexível para o registo de informação histórica no contexto da reconstituição de comunidades históricas.

3.2.3.3.Generalização e especialização de entidades históricas.

Ao generalizarmos o significado inicial das tabela "relações" abrimos a possibilidade de alargar o âmbito do esquema inicial da base de dados. Constituído por três tabelas básicas ("pessoas", "atributos", "relações"), este esquema visava aglutinar de modo flexível informação variável no tempo sobre pessoas, com o registo da fonte original onde cada item de informação foi recolhido. Contudo, como vimos na secção anterior, faz todo o sentido alargar o âmbito de aplicação da tabela de relações para incluir ligações não só entre pessoas mas também entre pessoas e outro tipo de entidades. Esta modificação

menor provoca uma flexão importante de filosofia cujas consequências analisaremos agora.

Re-examinemos a coluna "destino" da tabela "relações" na forma que assume no quadro 3.12. Temos dois valores distintos nessa coluna: "d1692-t1" e "dev1692", que ocorre duas vezes. O primeiro valor é a matrícula de uma pessoa, o segundo é a matrícula de um acto. Temos assim entidades de tipo diferente referidas numa mesma coluna. Essa variedade tem muitas vantagens, como já referimos, mas pode também causar alguns problemas técnicos e formais. Quando um programa informático analisa as relações de uma pessoa a partir desta tabela encontrará matrículas identificadoras de entidades muito diferentes. Para obter mais informação sobre "d1692-t1" terá que consultar a tabela das pessoas, para obter mais informação sobre "dev1692" terá que consultar a tabela de actos. De modo que o programa, para cumprir eficazmente a sua função, terá que poder determinar rapidamente que tipo de entidade corresponde a cada matrícula para poder completar a informação relevante. Se generalizarmos também a tabela de atributos teremos uma questão semelhante: onde agora temos uma coluna com a matrícula da pessoa a que o atributo pertence passaremos a ter matrículas correspondendo a actos, ou outro tipo de entidades.

É assim necessário prever um modo eficaz de determinar o tipo de entidade a que se refere uma matrícula de identificação qualquer. Existem vários métodos possíveis para levar a efeito essa tarefa. O que iremos propôr de seguida baseia-se no modelo entidade - relacionamento extendido que referimos anteriormente (3.2.2.1) e constitui um método testado de representar o conceito de generalização e especialização em sistemas de base de dados relacionais.

O primeiro passo a dar neste contexto é formular a noção de "entidade histórica" a um nível abstracto. Uma "entidade histórica" ou, se se quiser, "um objecto histórico", é tudo aquilo que pode ter uma matrícula identificadora e assim ser origem/destino de uma relação ou ainda possuir um número determinado de atributos. Chegamos ao conceito de "entidade histórica" por generalização das duas entidades que já tínhamos definido: "pessoas" e "actos". Como vimos, no nosso exemplo, tanto "pessoas" como "actos" aparecem na coluna "destino" da tabela de relações. Quase só com um artifício de redacção passamos a dizer que as matrículas que aparecem nessa coluna designam "entidades históricas" que podem ser ou "pessoas", ou "actos", ou outras entidades que futuramente queiramos representar.

Dentro deste quadro, as categorias "pessoa" e "acto" são casos particulares de "entidades históricas", ou, na linguagem do modelo, "especializações". Para representar esta informação adicional introduzimos uma nova tabela, com apenas duas colunas, denominada "entidades" (quadro 3.13).

Quadro 3.13: Tabela para registo de entidades

Entidades

Atributos/colunas	*Explicação*
id	matrícula identificadora da entidade
tipo	tipo de entidade (pessoa, acto, escritura,...)

A função dessa tabela é permitir aos programas que interagem com a base de dados determinar rapidamente que tipo de entidade corresponde a determinada matrícula. Preenchendo esta tabela com os dados do nosso

exemplo obtemos uma lista com três entidades, duas pessoas e um acto (quadro 3.14).

Quadro 3.14. Linhas na tabela de entidades

Entidades

Id	*Tipo*
d1692-r1	pessoa
d1692-t1	pessoa
dev1692	acto

Cruzando a informação da tabela de relações com a tabela de entidades qualquer programa obtém rapidamente a informação sobre o tipo de entidades envolvidas numa relação concreta. As colunas "origem" e "destino" da tabela de relações referem matrículas que necessariamente existem na coluna "id" da tabela "entidades". Essas matrículas podem corresponder a pessoas ou não. Do mesmo modo alteraremos a tabela "atributos" para a tornar geral e não apenas aplicável a atributos pessoais. A modificação consiste apenas em modificar o nome da coluna "pessoa" para a etiqueta mais genérica de "entidade" (ver quadro 3.15).

Quadro 3.15 Tabela para registo de atributos de entidades (II)

Atributos

Atributos/colunas	*Explicação*
id	matrícula identificadora (automática)
entidade	*id da entidade respectiva (refere a coluna "id" da tabela "entidades")*
atributo	tipo de atributo

valor	valor do atributo
data	data do atributo
fonte	fonte original (código)
obs	observações gerais

Deste modo podemos adicionar atributos a qualquer tipo de entidades e não apenas a pessoas.

Reparemos mais de perto na relação entre a tabela "entidades"(quadro 3.14) e as tabelas "pessoas" e "actos" (quadros 3.5 e 3.11). Cada linha das tabelas de pessoas tem uma linha correspondente na tabela de entidades e o mesmo acontece com a tabela de actos. A coluna "id" da tabela de entidades reproduz as colunas "id" das tabelas de pessoas e actos. De facto, cada vez que introduzirmos uma nova pessoa ou um novo acto teremos que registar na tabela de "entidades" a matrícula respectiva e o tipo de entidade que se trata. Assim a tabela "entidades" funciona como um registo central de todos os "objectos" existentes na base de dados. Cada "objecto" terá sempre pelo menos duas entradas na base de dados: uma na tabela geral de todos os objectos (a tabela "entidades") e outra na tabela correspondente ao tipo de objecto em questão (a tabela "pessoas" ou a tabela "actos", dentro do nosso exemplo). A mesma matrícula em ambas as tabelas serve de elemento de ligação.

Do ponto de vista conceptual dizemos que "pessoas" e "actos" são especializações de "entidades", ou "sub-classes" de "entidades". Do ponto de vista da implementação cada relação de especialização implica duas tabelas com chaves comuns. Esta metodologia pode ser aproveitada de forma genérica e gerar uma base de dados em que os vários tipos de entidades estão hierarquicamente organizados.

Para expandir esta temática com um exemplo adicional vamos introduzir o conceito de "devassa" de forma mais concreta na base de dados. Como vimos anteriormente a devassa de 1692 foi introduzida na base de dados como um "acto" cujo único atributo foi ter ocorrido em determinada data. Contudo temos outros atributos que descrevem a devassa enquanto acto. Por exemplo, o local em que decorreu, o nome do visitador, qual o pároco que era residente no momento da devassa, etc. Para isso necessitamos de "especializar" a entidade "acto" em algo mais concreto que é a "devassa". Tal como a especialização de "entidade histórica" em "acto" envolveu duas tabelas com chaves comuns, a especialização de "acto" em "devassa" vai envolver também duas tabelas que partilham chaves: a tabela de "actos", que já existe, e uma nova tabela de devassas que terá mais colunas para receber os valores dos atributos específicos de uma devassa (ver quadros 3.16 e 3.17).

Quadro 3.16: Tabela para registo de devassas

Devassas

Atributos/colunas	*Explicação*
id	matrícula identificadora da devassa
local	local onde decorreu a devassa
visitador	nome do visitador
paroco	nome do pároco residente
obs	observações genéricas

Quadro 3.17 Linhas na tabela das devassas

Devassas

id	*local*	*visitador*	*pároco*	*obs*
dev1692	misericórdia	manuel joao, de-	luis alvares pin-	devassa em vá-

	sembargador da to relação episco-pal.	rios dias (10 de Abril a 19 de Abril)

Assim o "objecto" histórico que constitui a devassa de 1692 produz três linhas em três tabelas: entidades, actos e devassas. Essas três linhas reflectem as relações de especialização crescente: uma devassa é um caso particular de acto que por sua vez é um caso particular de "entidade" histórica.

Os atributos da devassa, enquanto objecto histórico, distribuem-se pelas três tabelas da seguinte maneira: o "tipo", por ser um atributo comum a todas as entidades históricas, fica registado na tabela de "entidades"; a data, por ser um atributo comum a todos os actos, vai para a tabela de actos; o local, visitador e pároco, por serem específicos da devassa, vão para a respectiva tabela. A matrícula de identificação está presente em todas as tabelas para garantir que a informação é mantida coerentemente ligada.

Para consumar esta nova estrutura temos que modificar o registo da devassa na tabela de entidades, introduzindo a nova informação que o objecto "dev1692" não é simplesmente um acto, é algo de mais concreto, uma devassa (ver quadro 3.18).

Quadro 3.18. Linhas na tabela de entidades
(II - após a especialização das devassas)

Entidades

Id	*Tipo*
d1692-r1	pessoa
d1692-t1	pessoa
dev1692	**devassa**

Podemos agora resumir o esquema que nos permite formalizar entidades históricas variadas numa hierarquia de categorias.

O primeiro conceito formal é o de "entidade" ou "objecto" histórico. A sua definição é a mesma que a definição de "entidade" no modelo "entidade-relação": algo que existiu e que pode ser descrito por um conjunto de atributos. Cada entidade ou objecto pertence a um "tipo" ou "classe": pessoas, actos, devassas.

Este conceito formal tem a sua implementação concreta na tabela de entidades. A tabela de entidades regista todas as entidades existentes. Cada entidade tem uma matrícula e pertence a um determinado tipo ou classe. A tabela tem duas colunas que registam essa informação: a matrícula e o tipo ou classe de cada entidade existente.

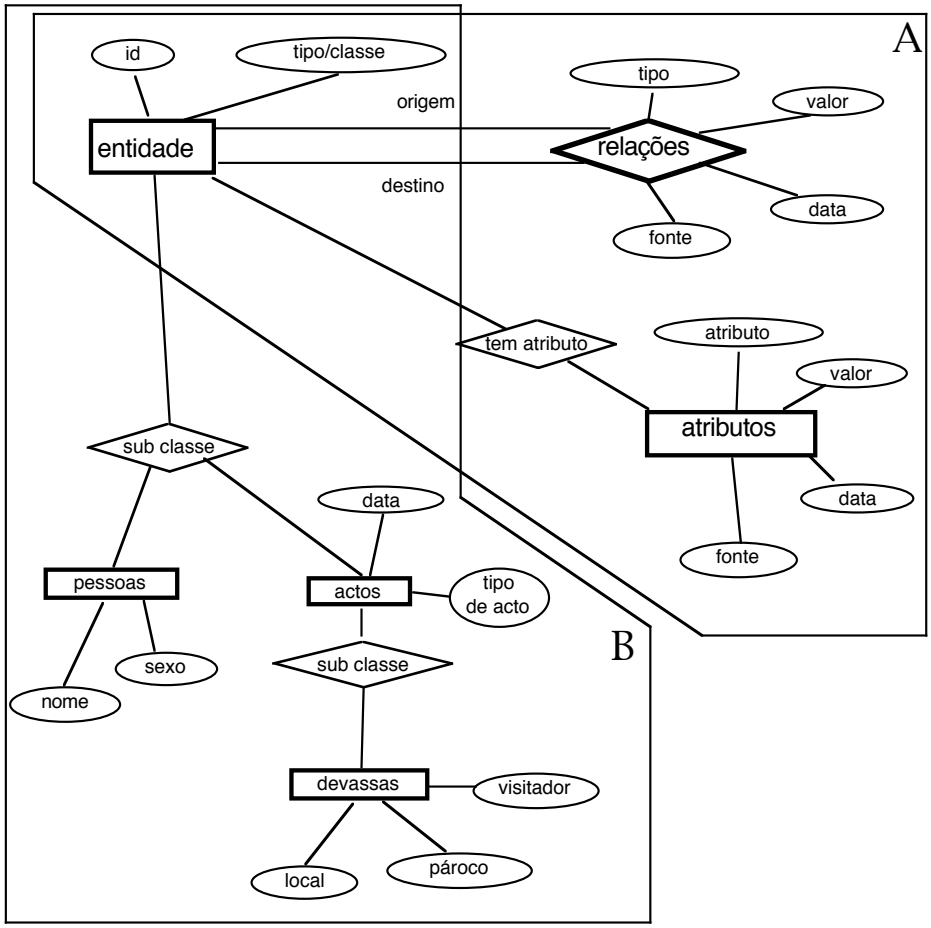
O segundo conceito formal é o de especialização. Mais uma vez nos baseamos no modelo entidade-relação na sua forma extendida. Uma especialização é a relação entre dois tipos de entidades em que um representa um caso particular, ou mais específico, do outro. Assim as devassas são uma especialização dos actos em geral. "Actos" e "pessoas" são especializações de "entidades em geral".

A implementação deste conceito passa pela criação de uma tabela para cada nível da hierarquia de especialização. A tabela "entidades" representa o nível mais alto, o de entidade histórica em geral. Para cada especialização cria-se uma tabela adicional que permite o registos dos atributos particulares do novo tipo de entidade. Quando inserimos um objecto histórico concreto, com o seu

conjunto de atributos fixos, distribuimos a informação pelas várias tabelas da hierarquia, conforme o tipo de entidade a que cada atributo corresponde.

Podemos representar a estrutura da base dados, tal como está definida até agora, por um diagrama (ver diagrama 3.1).

Diagrama 3.1 Estrutura nominal

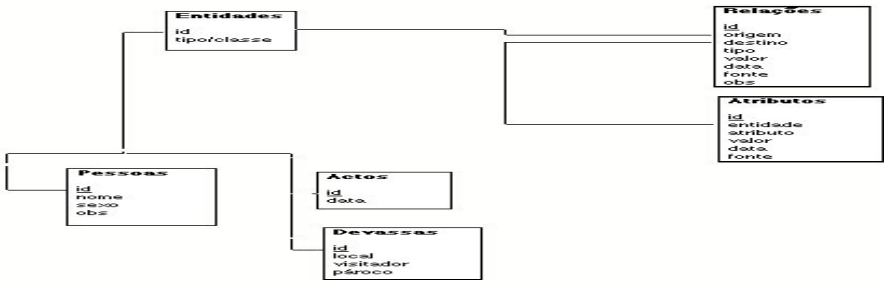


Este diagrama utiliza algumas convenções gráficas comumente aceites na representação de esquemas produzidos pelo modelo "entidade/relação". Os rectângulos indicam entidades e correspondem a tabelas na base de dados relacional. As ovais correspondem a atributos das entidades e às colunas das respectivas tabelas. Os losângulos correspondem a relações entre entidades.

Algumas relações possuem atributos associados, como é o caso do losângulo denominado "relações".

O diagrama tem duas zonas distintas. A primeira engloba a tríade entidades/relações/atributos, correspondendo a três tabelas na base de dados relacional (zona A). Este conjunto de estruturas tem a capacidade de representar objectos históricos genéricos com atributos variáveis e as respectivas relações, com a identificação da fonte em que cada item de informação originou. Esta tríade é estável e constitui o núcleo principal do formalismo. A segunda zona engloba os rectângulos "entidades", "pessoas", "actos" e "devassas" (zona B). É a zona onde se estruturam as relações hierárquicas de especialização. Esta zona vai captar a evolução futura do sistema. Aqui se adicionam novas categorias de objectos à medida que a informação a registar se vai complexificando. À representação abstracta da diagrama 3.1 podemos acrescentar um diagrama das tabelas envolvidas no esquema relacional (diagrama 3.2).

Diagrama 3.2 Estrutura de tabelas relacionais



O esquema formal básico fica deste modo definido. O resto dos elementos a introduzir corresponde a especializações progressivas na zona B do diagrama

da diagrama 3.1. mas não necessitam de conceitos radicalmente novos para além dos que foram explanados até aqui.

A próxima secção está ainda ligada à estrutura da base de dados e à representação formal da informação. Contudo foca já aspectos que estão mais directamente ligados a questões de implementação de um sistema específico, do que à lógica interna da informação histórica. Mais concretamente a secção seguinte explicita o modo como adicionamos à base de dados informação sobre a sua própria estrutura, permitindo que os programas actuem de maneira mais eficaz sobre a hierarquia de especializações. O principal objectivo do que aí se exporá é de mostrar como se podem implementar facilmente programas que actuam sobre os objectos históricos de modo diferenciado, conforme se trata de pessoas, actos, devassas, ou outro tipo. Para potenciar essa capacidade é necessário que a base de dados registe não só a informação histórica em si mas também informação mais abstracta sobre os vários tipos de entidades presentes nessa informação. O interesse desta matéria para o historiador é reduzido. O seu destinatário preferencial é a pessoa responsável por implementar um sistema concreto baseado nos formalismos explicados até agora. Introduziremos alguns conceitos e extensões à estrutura da base de dados que fazem sentido mais do ponto de vista da lógica da implementação informática do que da lógica da investigação histórica. Para quem quiser (ou tiver que) preocupar-se com esse aspecto a próxima secção introduz elementos que facilitarão bastante o processamento de uma base de dados relacional que contém informação hierarquicamente organizada.

3.2.4. INTRODUZINDO META-INFORMAÇÃO: OBJECTOS, CLASSES E MÉTODOS.

Na secção anterior vimos como se pode criar uma estrutura de dados que representa um conjunto de entidades que estabelecem entre si relações de especialização. Essa estrutura define uma hierarquia que vai do geral ao particular por etapas sucessivas. Chegámos a essa solução por imposição derivada das características da informação a registar. Procedemos por isso inferencialmente, partindo de exemplos concretos de dados até chegarmos a uma estrutura abstracta.

Agora iremos examinar algumas consequências dessa estrutura do ponto de vista da implementação de programas que sobre ela actuem. Uma questão já aflorada acima diz respeito à necessidade de criar mecanismos que actuem diferentemente sobre os vários tipos de entidades. Referimos essa necessidade quando, analisando as relações de uma pessoa, vimos que os destinatários dessas relações podiam ser quer outras pessoas, quer actos. Na altura salientámos que um programa que estivesse a processar essa lista de relações teria que "saber" que existe informação específica adicional para as pessoas e informação específica para os actos. Mais concretamente o programa necessita de determinar rapidamente, para uma matrícula de identificação qualquer, o tipo de entidade respectivo. A tabela de entidades, que associa cada uma das matrículas a um tipo de entidade, foi criada a partir dessa necessidade.

Contudo, a estrutura de dados, tal como foi descrita até agora, não contém alguma informação importante para que os programas actuem de modo inteligente. Por exemplo, o facto de o tipo de entidade "devassa" ser uma especialização de "actos" que por sua vez é uma especialização de "entidade", é algo que nós sabemos mas não está explicitado na base de dados. Os programas que actuem sobre estes dados têm de obter esse conhecimento de algum modo.

É má política de implementação contruir programas que assumem uma determinada estrutura muito específica, sobretudo quando é suposto que essa estrutura evolua rapidamente. No nosso esquema, tal como está representado nos diagramas 3.1 e 3.2, temos algumas características que são genéricas e estáveis (a zona A) e outras que evoluem ao sabor das investigações concretas (a zona B). A parte da estrutura que evolui mais imprevisivelmente é precisamente aquela que diz respeito à tipologia das entidades representadas e às suas relações hierárquicas. Cada nova fonte pode criar novos tipos de entidades, e mesmo uma re-análise de fontes já tratadas, pode criar novos objectos de interesse que necessitem de novos tipos de entidades no sistema. Se os programas que operam sobre os dados são feitos especialmente para uma determinada configuração da estrutura de especialização (zona B) isso significa que na prática será difícil alterar essa mesma estrutura, sem dispender um grande esforço de reprogramação.

O ideal seria que os programas fossem genéricos e se adaptassem automaticamente às evoluções da zona B da diagrama 3.1. Mas para que isso seja possível é necessário que a nossa base de dados, além de fornecer aos programas a informação originária das fontes histórica, forneça também a informação que descreve a sua organização interna. Por outras palavras, a base de dados tem de encerrar, sob a forma de dados, a descrição da sua própria estrutura. Desse modo podemos imaginar programas que interrogam a base de dados, num primeiro momento, para obterem uma descrição das entidades existentes e respectivas relações de especialização e, num segundo momento, actuam sobre a informação histórica referente a essas entidades. Conseguiríamos, por este processo, programas genéricos que se auto-adaptariam às características específicas de determinados contextos de investigação.

A criação de programas que reagem diferencialmente conforme a natureza do objecto a que se aplicam é um dos temas centrais da investigação em técnicas de programação. Chama-se "programação orientada por objectos" à metodologia utilizada para construir programas com as características referidas. Como se trata de uma matéria consagrada, com ampla literatura, procuraremos utilizar o seu vocabulário específico que, aliás, retoma com termos diferentes os principais conceitos até agora introduzidos.

Em programação orientada por objectos, os "tipos de entidades" que temos vindo a referir designam-se por "classes". Temos assim, nos exemplos que temos vindo a propôr, as classes "entidade","pessoa" e "acto". Denominam-se "objectos" as instâncias concretas dessas classes, por exemplo a pessoa "d1692-t1" ou o acto "dev1692". As classes têm determinados atributos, em tudo conceptualmente idênticos aos atributos elementares das entidades do modelo ER. Cada objecto de uma classe possui valores próprios para esses atributos comuns a todos os objectos da mesma classe. As classes organizam-se hierarquicamente através de relações de especialização iguais às que descrevemos para as nossas entidades. Quando uma classe especializa outra chama-se à classe mais geral a "super-classe" e à mais específica a "sub-classe". Assim diríamos que "devassa" é uma sub-classe de "acto", ou que, inversamente acto é a "super-classe" de devassa.

Os procedimentos, ou programas, que actuam sobre os objectos, estão associados às classes respectivas. Assim determinadas funções básicas como "imprimir", "criar", "apagar", "alterar" são programas para uma determinada classe. Quando o utilizador dá uma instrução concreta, como por exemplo "imprimir" um determinado objecto, o sistema em primeiro lugar determina a classe a que o objecto pertence. Seguidamente determina se existe algum programa de impressão específico para essa classe. Se existe executa-o e a tarefa está determinada. Se não existe então o sistema procurará aplicar uma solução

mais geral: irá determinar se a classe do objecto em causa é uma especialização de uma classe mais genérica. Se assim fôr procurará de novo se existe um programa de impressão para a classe mais geral e executa-o nesse caso. O sistema continuará pela hierarquia acima até chegar ao topo.

Através deste mecanismo a funcionalidade criada a determinado nível da hierarquia propaga-se pelos níveis inferiores. Em programação orientada por objectos denomina-se "herança" a este mecanismo de delegação de funcionalidade das classes mais gerais nas classes menos gerais. Quando se procura efectuar uma determinada operação com um objecto, como por exemplo "imprimir", diz-se que se envia a mensagem "imprimir" a esse objecto. As classes associam programas a mensagens. Um programa que dá seguimento a uma mensagem chama-se um "método".

Um exemplo talvez ajude a clarear os conceitos.

Suponhamos que se define um programa para imprimir um objecto da classe "entidade", que corresponde ao nível mais alto da nossa hierarquia. As entidades, recordemos, só possuem dois atributos: id e tipo. O programa especificaria basicamente algo do género:

```
Para imprimir uma entidade fazer (método p1): tipo$id
      em que id  e tipo  serão substituídos pelos valores reais.
```

Se ordenarmos ao sistema para imprimir os três objectos do nosso exemplo (ver quadro 3.18) seria produzido o seguinte resultado:

```
pessoa$d1692-t1
pessoa$d1692-r1
devassa$dev1692
```

O que acontece é que o sistema em primeiro lugar determina a classe de cada objecto e procurará um método de impressão específico. Como não encontra nenhum, a mensagem "imprimir" sobe na hierarquia para a classe "entidade" onde encontra o método mais geral, aquele que funciona sempre. No caso do objecto "dev1692" primeiro procura um método associado à classe "devassa" e falha. Depois procura se existe na classe "acto" e falha novamente. Finalmente utiliza o método para a classe "entidade".

Seguidamente definiríamos a mesma função de impressão para a classe "acto":

```
Para imprimir um acto fazer (método p2): tipo de acto$id/data
```

A mesma mensagem de impressão enviada aos três objectos produziria agora um resultado diferente:

```
pessoa$d1692-t1
pessoa$d1692-r1
devassa$dev1692/16920410
```

Agora no caso de "dev1692" é utilizado o método da classe "acto". Obviamente poderíamos definir um programa específico para devassas:

```
Para imprimir devassas fazer  (método p3):
     devassa$data/id=id
        /visitador=visitador
        /paroco=paroco
        /obs=obs
```

O resultado agora seria:

```
pessoa$d1692-t1
pessoa$d1692-r1
devassa$16920410/id=dev1692
     /visitador=manuel joao, desembargador da relação episcopal.
```

```
/paroco=luis alvares pinto
/obs=devassa em vários dias (10 de Abril a 19 de Abril)
```

Note-se que outros actos que não devassas que eventualmente existissem continuariam a ser impressos na forma *tipo de acto*$*id*/*data*.

Vejamos agora o que é necessário para implementar um sistema com estas características sobre a base de dados descrita. A informação adicional necessária é de dois tipos: em primeiro lugar temos que registar as relações de especialização entre as classes; em segundo lugar temos de registar os métodos associados a determinadas mensagens, como "imprimir", para as classes em que estão definidas. A esta informação adicional que descreve as características abstractas da informação do mundo real chamaremos "meta-informação".

Para manter a coerência formal do sistema esta nova informação será registada como dois novos tipos de entidades, que especializam a entidade genérica. Necessitaremos pois de dois tipos de entidades adicionais. A entidade "classes" registará a informação relativa aos tipos de entidades. A tabela "métodos" associará métodos a classes para o tratamento de mensagem específicas. Estas novas entidades serão traduzidas, pelo mesmo processo que usámos anteriormente, por duas novas tabelas: "classes" e "métodos". Cada linha destas novas tabelas será individualizada por uma matrícula identificadora e gerará uma nova entrada na tabela geral de entidades. Devido ao limitado número de classes que vamos definir não se justifica utilizar matrículas automáticas para identificar as classes. Usaremos o próprio nome da classe como identificador. Para cada classe registamos o nome da classe imediatamente superior na hierarquia e o nome da tabela do esquema relacional que irá receber a informação dos objectos referentes a esta classe.

A tabela de classes tem a estrutura mostrada no quadro 3.19.

Quadro 3.19 Tabela para registo de classes

Classes

Atributos/colunas	*Explicação*
id	nome da classe
super	nome da super-classe desta classe
tabela	tabela que regista os objectos desta classe

Esta tabela receberá as linhas referentes aos tipos de entidades já definidos nos exemplos da secção anterior: "entidade", "pessoa","acto", "devassa", a que teremos que acrescentar os dois novos tipos agora definidos, "classe" e "método".

Quadro 3.20: Linhas na tabela de classes

Classes

id	*super*	*tab ela*
entidade		entidades
pessoa	entidade	pessoas
acto	entidade	actos
devassa	acto	devassas
classe	**entidade**	**classes**
metodo	**entidade**	**metodos**

Esta tabela descreve a zona B do diagrama 3.1. Repare-se que a classe entidade não tem super classe, uma vez que está no topo da hierarquia. Todas as outras classes são especializações de "entidade", excepto "devassa" que especializa a classe "acto".

Um factor importante da informação contida na tabela "Classes" consiste no enquadramento abstracto que é dado às tabelas do esquema relacional. Assim as tabelas passam a estar associadas hierarquicamente uma vez que a cada tabela corresponde uma classe devidamente posicionada na hierarquia de especialização. As únicas tabelas do esquema relacional que não estão referidas na tabela de "classes" são as associadas às relações variáveis e aos atributos variáveis (zona A da figura 3.1). Com uma generalização adicional, a última de todo este processo, iremos considerar as relações variáveis e os atributos variáveis como entidades pertencendo a classes próprias. Vamos por isso incluir mais duas classes na tabela do quadro 3.20 e assim obter uma descrição completamente auto-contida da base de dados (ver quadro 3.21).

Quadro 3.21: Linhas na tabela de classes (II)
(Incluido relações e atributos)

Classes

id	*super*	*tabela*
entidade		entidades
pessoa	entidade	pessoas
acto	entidade	actos
devassa	acto	devassas
classe	entidade	classes
método	entidade	métodos
relação	**entidade**	**relações**
atributo	**entidade**	**atributos**

Os diferentes programas existentes são registados na tabela "métodos". Cada linha dessa tabela associa três elementos: um método, identificado por

uma matrícula, a classe a que está associado e a mensagem que o activa. Os quadro 3.22 e 3.23 descrevem essa tabela e exemplificam com os três programas de impressão imaginados acima.

Quadro 3.22 Tabela para registo de métodos

Métodos

Atributos/colunas	*Explicação*
id	nome do método
classe	nome da classe a que está associado
mensagem	mensagem que activa este método

Quadro 3.23 Linhas na tabela de métodos

Método

id	*classe*	*mensagem*
p1	entidade	imprimir
p2	acto	imprimir
p3	devassa	imprimir

A tabela de entidades passa a reter os novos objectos que encapsulam a meta-informação:

Quadro 3.24 Linhas na tabela de entidades
(III - após a introdução de meta-informação e considerando "relações" e "atributos")

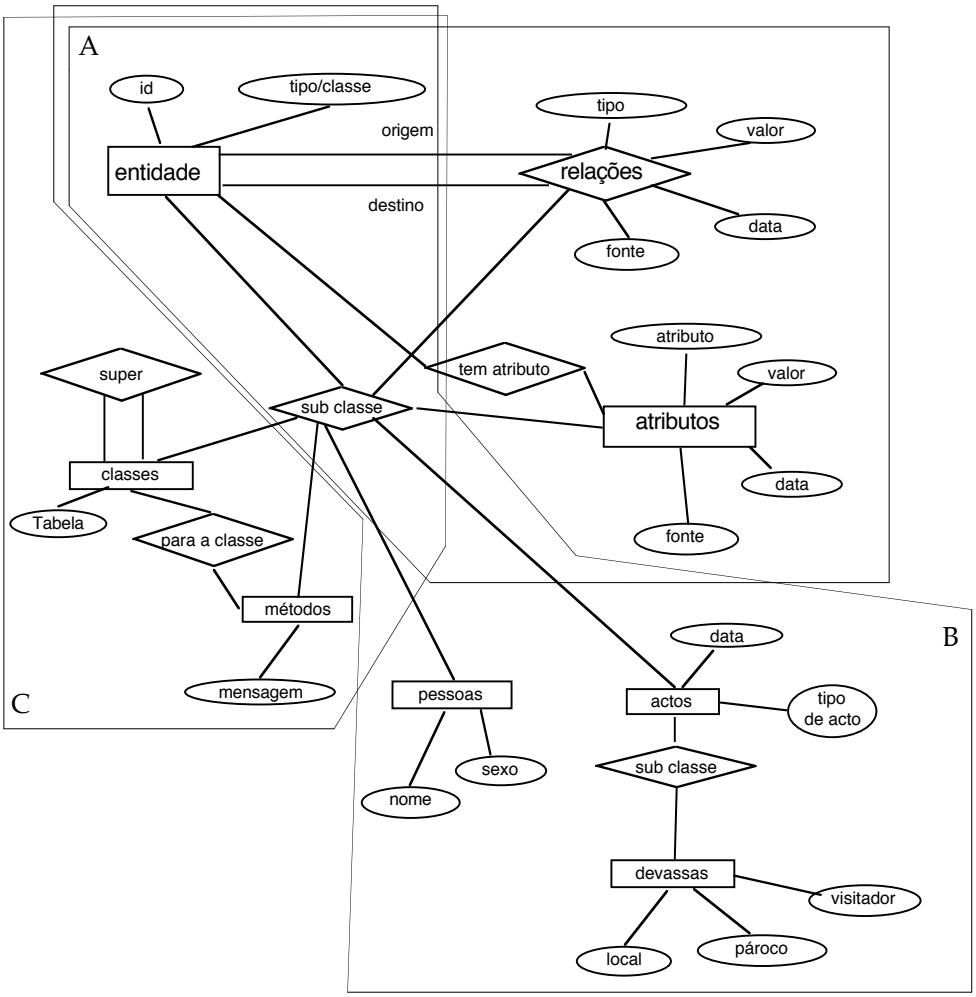
Entidades

Id	*Tipo*
d1692-r1	pessoa
d1692-t1	pessoa

dev1692	devassa
d1692-r1r1	**relação**
d1692-t1r1	**relação**
d1692-r1r2	**relação**
d1692-t1a1	**atributo**
d1692-t1a2	**atributo**
d1692-t1a3	**atributo**
d1692-r1a1	**atributo**
entidade	**classe**
pessoa	**classe**
acto	**classe**
devassa	**classe**
classe	**classe**
metodo	**classe**
p1	**método**
p2	**método**
p3	**método**

Podemos agora rever o esquema 3.1 para incluir a meta-informação sobre classes e métodos (ver diagrama 3.3).

Diagrama 3.3: Estrutura com meta-informação



No novo esquema passam a existir três zonas. As zonas A e B mantêm-se como anteriormente e a nova zona C encerra a meta-informação que descreve o resto da base de dados, incluindo os programas disponíveis. Repare-se também que adicionámos linhas a especificar uma relação de subclasse entre "relações" e "atributos", por um lado, e "entidades" por outro. Deste modo todas as tabelas da base de dados correspondem a uma classe de entidades que descende por especialização da classe "entidade".

Esta estrutura realiza os objectivos apontados para a reconstituição de comunidades históricas. Permite descrever objectos históricos definidos por

atributos variáveis no tempo e interligados por relações que variam igualmente no tempo. Cada atributo e cada relação são retraçáveis até às fontes originais. A base de dados organiza os vários tipos de objectos numa hierarquia de classes. Essa hierarquia faz parte da própria base de dados que, de certo modo, se auto descreve. A capacidade auto-descritiva da base de dados permite uma estruturação de programas com um sabor "orientado a objectos", que aumenta a flexibilidade e a facilidade de expansão do sistema.

A possibilidade de explorar esta estrutura de forma eficaz, em termos de implementação concreta, depende muito das ferramentas de programação utilizadas. Certas soluções permitirão explorar mais eficazmente a meta-informação disponível. Outras deixarão pouca margem para a produção de programas verdadeiramente flexíveis. Os apêndices desta parte fornecerão mais detalhes sobre as ferramentas utilizadas, as vantagens e limitações encontradas, e o modo como soluções alternativas poderiam aumentar a funcionalidade.

Os conceitos aqui apresentados têm, contudo, algum valor autónomo. Descrevem de modo o mais simples que foi possível os tipos de conceitos, problemas, e soluções envolvidos na análise da representação de dados com as características definidas no início. Existe, em consequência, um aspecto formal que é prévio às "contaminações" da implementação e das cedências às particularidades das ferramentas de desenvolvimento. O essencial desta proposta, do ponto de vista das metodologias de base de dados, reside no sistema de implementar uma hierarquia ER extendida com características de orientação por objectos dentro de uma base de dados relacional. Em apêndice forneceremos uma descrição mais formalizada desse sistema, das regras que lhe estão subjacentes e dos detalhes da sua implementação.

3.3. A REPRESENTAÇÃO DE DOCUMENTOS E O REGISTO DE DADOS.

3.3.1. DA FONTE À BASE DADOS: O PROBLEMA

No capítulo anterior desenvolvemos uma estrutura flexível de base de dados para suportar a informação recolhida durante uma investigação nominal. A análise feita procurou criar um esquema versátil do ponto de vista da variedade de informação a receber e dos processos a executar. O processo de refinação progressiva dessa estrutura foi sempre guiado por preocupações estritamente formais, sem considerar o modo como, na prática, o historiador ou o operador procederia à efectiva transcrição dos documentos históricos para o sistema.

O resultado da análise formal deixa-nos uma estrutura que, pelo seu grau de abstração, está muito afastada do modo como a informação histórica chega até nós. Uma maneira de realçar este problema é comparar o fragmento

documental que desencadeou a nossa análise com a informação depois de formalizada.

Pondo de parte a meta-informação descrita no fim da secção anterior e restringindo-nos puramente à informação histórica, o fragmento inicial, que consistia na frase:

> José Machado, solteiro, da vila, que de idade disse ter 24 anos, *filho* de Manuel Fernandes o ratinho....

Quadro 3.25: Linhas na tabela de pessoas (repetição)

Pessoas

id	*nome*	*sexo*	*obs*
d1692-t1	jose machado	m	
d1692-r1	manuel fernandes	m	

Quadro 3.26: Linhas na tabela de atributos (repetição)

Atributos

id	*pessoa*	*atributo*	*valor*	*data*	*fonte*	*obs*
d1692-t1a1	d1692-t1	ec	s	16921004	ldp1692	
d1692-t1a2	d1692-t1	residencia	soure	16921004	ldp1692	"vila"
d1692-t1a3	d1692-t1	idade	24	16921004	ldp1692	
d1692-r1a1	d1692-r1	alcunha	ratinho	16921004	ldp1692	

Quadro 3.27 Linhas na tabela de relações (iepetição)

Relações

id	*origem*	*destino*	*tipo*	*valor*	*data*	*fonte*	*obs*
d1692-r1r1	d1692-r1	d1692-t1	parentesco	pai	16921004	ldp1692	
d1692-t1r1	d1692-t1	dev1692	funcao em acto	teste-munha	16921004	ldp1692	

Esta transformação operada sobre a informação original no sentido de a enquadrar na estrutura formalizada de dados é ao mesmo tempo necessária e fundamental. É por esta operação que a informação histórica se torna verdadeiramente manipulável. Introduzimos assim uma distância entre a fonte e a base de dados que, se se justifica em termos formais, coloca problemas de utilização prática.

Existe sempre uma tensão intrínseca entre as necessidades de formalização da informação e a necessidade de manter processos de registo de dados próximos das fontes. Estruturas próximas das fontes normalmente facilitam a introdução de dados mas dificultam o processamento posterior. Inversamente, estruturas de dados criadas de modo a optimizar a fase de exploração afastam-se em regra da organização da fonte e dificultam o processo de registo. Esta contradição só se esbate quando a estrutura da fonte é muito simples e regular aproximando-se das necessidades de normalização postas pela fase de processamento de dados.

Ao olharmos o resultado do tratamento do fragmento de fonte que serviu o nosso exemplo apercebemo-nos que a introdução de dados exige um trabalho de "normalização" da informação que consiste sobretudo na separação dos vários elementos pelas colunas das várias tabelas interligadas. Se esse trabalho tiver que ser executado pelo operador ou pelo historiador o custo de informatização pode tornar-se excessivo e o processo de introdução de dados demasiado longo e sujeito a erros. Experiências feitas demonstraram que mesmo com programas que permitissem abrir várias janelas para introdução simultânea de dados em várias tabelas, o esforço de registo de fontes dentro deste esquema formal era excessivo e altamente sujeito a erros.

A solução encontrada para este problema foi de automatizar o processo de "normalização" das fontes[20]. Criaram-se assim uma série de programas que processavam um texto que continha a informação registada de um modo muito próximo das fontes originais, produzindo automaticamente e sem intervenção humana os valores a introduzir nas várias tabelas. Com estes programas o processo de registo das fontes deixou de recorrer aos sistemas de gestão de base de dados, que incluem os seus próprios métodos de introdução de informação, para utilizar como ferramenta principal um processador de texto simples. As fontes são transcritas como textos, com uma forma legível e envolvendo um certo número de convenções de escrita fáceis de apreender. Esses textos são posteriormente submetidos aos programas de tradução que efectuam a normalização produzindo a informação organizada em função das várias tabelas do esquema relacional. Uma fase final, denominada importação, incorpora na base de dados o resultado da tradução.

Esta abordagem tem alguns problemas próprios, que veremos mais adiante, mas resolve a dificuldade essencial: a da distância criada entre a fonte e a estrutura formal final. Ao automatizar o processo de normalização permitiu que a concepção do esquema da base de dados fosse determinado principalmente por uma lógica de correcção formal e não por necessidades de rentabilização do registo de dados. Permitiu igualmente uma certa independência do registo de dados em relação às variações que o esquema relacional foi sofrendo. Com efeito, ao longo do desenvolvimento do sistema, a estrutura das tabelas foi-se apurando e racionalizando sem que isso tivesse qualquer consequência sobre as fontes já registadas nem sobre as rotinas de trabalho estabelecidas. Modificações estruturais são facilmente incorporadas

---

[20] Ver, para uma abordagem diferente da que expomos aqui: Breure, Leen, *Interactive Data Entry: Problems, Models, Solutions*, "History and Computing", vol. 7, n.1, 1995, p.30-49.

nos programas de tradução que releêm os textos originais e produzem os dados re-organizados sob um novo modelo.

A implementação desta solução tem duas componentes principais: uma é a definição das regras de transcrição que os textos produzidos a partir das fontes devem seguir; a segunda é a criação de programas informáticos capazes de ler esses textos e produzir o resultado "normalizado" para inserção automática na base de dados. As secções seguintes fornecem mais detalhes sobre este dois aspectos, explicitando ainda o funcionamento dos tradutores e os problemas relativos à introdução de dados que ainda não foram eficazmente resolvidos.

3.3.2. A LINGUAGEM DE DESCRIÇÃO DE DOCUMENTOS

Sendo as fontes históricas no âmbito da reconstituição de comunidades históricas essencialmente textos, a forma de registo de dados mais próxima dos documentos é igualmente um texto. A utilização de fichas, neste tipo de investigação, não se afigura muito eficaz. De facto, a necessidade de recolher informação nominal muito variada coexiste dificilmente com a rigidez dos formulários de entrada de dados, quer em papel, quer directamente em computador. O texto, sendo naturalmente extensível e podendo seguir a sequenciação de elementos da própria fonte revela-se a forma ideal de transcrição do documento.

Contudo as tecnologias existentes estão muito longe de poderem processar uma transcrição *ipsis verbis* de uma fonte histórica. Algum grau de uniformização e anotação é imprescindível para permitir um processamento automático dos textos produzidos. Essa uniformização assume a forma de uma "linguagem" particular a que os textos de registo devem obedecer. Trata-se de

um linguagem extremamente simples mas que possui, não obstante, uma gramática e uma semântica próprias.

Apesar de na elaboração deste sistema se terem produzido algumas especificações de notações ou linguagens para este efeito a escolha final acabou por recair num sub-conjunto da linguagem de transcrição definida pelo programa Kleio, do Max-Plank Intitute für Geschichte[21]. Apesar do sistema Kleio utilizar formalismos próprios para a representação da informação na sua base de dados interna, a linguagem que propõe para o registo de documentos tem um valor intrínseco e pode ser utilizada indepentemente do resto do sistema. O próprio autor do sistema, Manfred Thaller, propôs uma notação universal para o registo de fontes históricas baseado no modelo da linguagem Kleio. Do nosso ponto de vista, a linguagem é simples de entender, económica na sua grafia e perfeitamente adequada às necessidades das fontes que utilizámos. Não deve contudo ser esquecido que estamos a utilizar simplesmente a notação de Kleio como mecanismo de entrada de dados e que o formalismo final desse sistema é de natureza muito diferente. A questão da equivalência entre a estrutura de um documento transcrito em Kleio e o modelo de dados utilizado aqui será detalhada nas páginas seguintes.

---

21 Thaller, Manfred, *A draft proposal for a standard for the coding of Machine Readable Sources*. "Historical Social Research / Historische Sozial forchung" 40, (1986), p. 3-46, onde é feita uma análise detalhada dos vários componentes de descrição de dados históricos; *Kleio: a database system*. Gottingen: Max-Planck-Institut, 1993. O formato de transcrição de dados usado pelo sistema Kleio é especificado nesta última obra p.13-37.

3.3.3. OS COMPONENTES DA LINGUAGEM KLEIO: DOCUMENTOS, GRUPOS, ELEMENTOS E ASPECTOS.

A linguagem Kleio baseia-se num certo número de conceitos que organizam os vários tipos de informação. Esses conceitos têm muitas semelhanças com os que vimos no capítulo anterior a propósito do modelo entidade-relação. Tratando-se em ambos os casos de modelos de formalização da informação é natural que exista uma certa sobreposição conceptual. Na exposição que se segue faremos sempre que fôr pertinente remissivas para os conceitos já expostos, assinalando as semelhanças e diferenças.

Kleio utiliza dois conceitos - "grupo" e "elemento" - com um significado muito semelhante ao de "entidade" e "atributo"[22]. Um "grupo" é a unidade principal de recolha de informação e corresponde a uma pessoa ou uma propriedade, por exemplo. O "grupo" é descrito por "elementos" que recebem valores concretos, do mesmo modo que os atributos do modelo entidade-relação. Kleio introduz ainda um nível abaixo do elementos ou atributos, definindo que cada valor de um elemento possa ter três "aspectos" ou formas: forma "base", forma "original" e "comentário". A forma "base" corresponde ao valor que normalmente se introduz num campo de um ficheiro. A "forma original" diz respeito ao modo como a informação estava de facto registada no documento, na sua expressão literal. O comentário é uma nota que o investigador regista no momento da transcrição e que se manterá associada ao elemento. Na prática os "aspectos" em Kleio servem para registar de forma rápida e eficaz informação adicional[23].

Finalmente em Kleio existe o conceito que um grupo pode incluir vários outros grupos. Por exemplo, a informação de um baptismo é registada num

---

[22] Thaller, Manfred, *Kleio...., cit.,* p. 18-27.

[23] Thaller, Manfred, *op. cit.,* p. 13.

grupo "baptismo" que inclui outros grupos como "pai", "mae", "padrinho", etc... O grupo de nível superior para uma determinada fonte designa-se por "documento". Um ficheiro de transcrição de uma fonte em Kleio é assim um ficheiro de texto composto por uma sequência de documentos com os respectivos elementos e sub-grupos.

Este sistema de três níveis (grupos, elementos e aspectos) transcreve-se numa notação extremamente simples.

O nome dos grupos deve ser seguido do sinal "$" e deve ser a primeira palavra de uma linha. Os elementos são registados sob a forma "nome"="valor". Elementos sucessivos são separados por uma barra ("/"). Para acelerar o registo é possível pré-definir a ordem dos elementos e omitir os nomes. Os "aspectos" são precedidos dos sinais "%" (linguagem original) e comentário "#".

Retomemos o exemplo da devassa referida na secção anterior. Cada devassa é uma entidade descrita pelos seguintes atributos: id, data, local, visitador, paroco, obs. Em Kleio temos um grupo "devassa" e os elementos respectivos. Poderíamos assim registar esta informação com a notação de Kleio, criando um texto com o seguinte conteúdo:

```
devassa$id=dev1692/data=16921004/visitador=manuel  joao
      /paroco=luis alvares pinto
                        /obs=devassa em vários dias
```

O nome dos elementos pode ser omitido desde que estes sejam introduzidos sempre pela mesma ordem. Assim o registo simplifica-se:

```
devassa$dev1692/16921004/manuel joao/luis alvares pinto/devassa em
vários dias
```

Cada pessoa é registada como um novo grupo, dependente deste e composto por grupos adicionais correspondendo aos atributos e relações:

```
pessoa$dev1692-t1/jose machado/m
      atr$ec/s
      atr$residencia/soure%vila
      atr$idade/24
pessoa$dev1692-r1/manuel fernandes/m
      atr$alcunha/ratinho
      rel$parentesco/pai/jose machado/dev692-t1
```

Como veremos de seguida este esquema pode ainda ser bastante simplificado para permitir uma transcrição rápida e próxima do original.

Compete aos programas de tradução ler estas descrições e gerar as linhas necessárias para preencher as tabelas do modelo relacional. Uma vez que controlamos o processo de tradução podemos incluir nos tradutores vários mecanismos de simplificação do registo. Como se pode ver no exemplo acima, não é necessário incluir a data de cada atributo, nem a respectiva fonte. O tradutor utilizará a data da devassa para todos os atributos pessoais registados. O mesmo se passa com o registo da fonte donde cada atributo provém.

Por este processo eliminamos a segmentação subjacente ao modelo relacional proposto e conseguimos um equilíbrio entre as necessidades de registo de dados e as imposições formais do modelo relacional.

Na secção seguinte veremos mais de perto como se faz a interacção entre os tradutores e o modelo relacional.

3.3.4. UTILIZAÇÃO DA LINGUAGEM KLEIO NO CONTEXTO DO SISTEMA RELACIONAL

3.3.4.1.Registo de pessoas

Comecemos pela descrição do registo da informação sobre as pessoas.

A forma genérica de registar uma pessoa é a seguinte:

```
pessoa$josé machado/m
```

Este item produzirá uma linha na tabela de pessoas. O "id" dessa pessoa será automaticamente criado no processo de tradução. Evitamos obrigar os utilizadores a inventarem constantemente matrículas unívocas para as pessoas, a não ser quando isso é estritamente necessário. O utilizador pode adicionar um elemento "obs" com observações genéricas:

```
pessoa$josé machado/m/obs=ja' apareceu em devassas anteriores.
```

Os vários atributos da pessoa são introduzidos imediatamente a seguir:

```
pessoa$josé machado/m/obs=ja' apareceu em devassas anteriores.
      atr$ec/s
      atr$residencia/soure%vila
      atr$idade/24
```

Cada atributo ocupa uma linha (ou mais caso o texto a introduzir o justifique) começada pelo grupo "atr$". Seguem-se o nome do atributo e o respectivo valor. Como foi referido os outros elementos da tabela dos atributos (fonte e data) são automaticamente preenchidos pelo tradutor a partir dos dados do documento corrente.

O registo de relações entre as pessoas envolve um mecanismo um pouco mais complexo. Ao registarmos uma relação temos de nomear a pessoa destino de modo não ambíguo. Para isso necessitamos de entrar em linha de conta com a matrícula de identificação dessa pessoa. Como os tradutores, por princípio, geram automaticamente as matrículas de identificação durante o processo de tradução, essas matrículas não são conhecidas no momento do registo. Assim, quando registamos uma pessoa que temos de referir noutro ponto de um documento é necessário atribuir-lhe uma matrícula directamente, desactivando a geração automática de "id"s pelo tradutor. Chamamos a isto a *atribuição explícita de uma matrícula* a uma pessoa, ou a qualquer outro objecto tratado pelo sistema.

Assim o registo que José Machado é filho de António Fernandes o Ratinho fica do seguinte modo:

```
pessoa$josé machado/m/obs=ja' apareceu em devassas anteriores.
      atr$ec/s
      atr$residencia/soure%vila
      atr$idade/24
      rel$parentesco/filho/manuel fernandes/dev692-r1
pessoa$manuel fernandes/m/id=dev692-r1
      atr$alcunha/ratinho
```

O elemento "id" no registo de uma pessoa indica ao tradutor que deve utilizar uma matrícula específica em vez de gerar uma automaticamente. Este processo torna possível a referência de pessoas sem ambiguidades, não obrigando contudo o utilizador a "inventar" matrículas senão quando estritamente necessário.

Note-se ainda o modo como se regista a relação dentro do grupo "rel". Introduzimos quatro elementos: o tipo da relação, o valor, o nome da pessoa destino e o seu "id". Em rigor não seria necessário indicar o nome da pessoa destino uma vez que a matrícula é suficiente para identificar sem

ambiguidades uma ocorrência individual. A presença do nome serve simplesmente para adicionar legibilidade à transcrição e permitir ao tradutor uma verificação adicional. Com efeito os tradutores são programados para verificar se as pessoas cujas matrículas são introduzidas em registo de relações têm de facto os nomes que são indicados. Este nível de redundância e de verificação de erros tornou-se bastante útil no processo de registo.

A informação funcional sobre as pessoas é registada de forma bastante económica. Em vez de registarmos todas as pessoas sob o grupo genérico "pessoa" utilizamos palavras diferentes conforme as funções que as pessoas cumprem no documento. Assim, sendo José Machado uma testemunha em devassa, usamos o termo "testemunha" em vez de "pessoa". Já Manuel Fernandes, pai de José Machado, não cumpre nenhuma função directamente relacionada com a devassa e por isso é registado como "referido". Ao utilizarmos a função da pessoa como início do registo da informação individual tornamos a transcrição do documento muito legível e compreensível mesmo para quem não tenha tido acesso a demoradas explicações sobre o formalismo utilizado. Por outro lado o nome da função permite muitas vezes inferir o sexo da pessoa (referido/referida, noivo/noiva,etc...). Poupa-se assim o registo explícito do sexo, que é sempre fonte de possíveis erros. O nosso exemplo fica então com a seguinte forma:

```
testemunha$jose machado/m/obs=ja' apareceu em devassas anteriores.
      atr$ec/s
      atr$residencia/soure%vila
      atr$idade/24
      rel$parentesco/filho/manuel fernandes/dev692-r1
referido$manuel fernandes/id=dev692-r1
      atr$alcunha/ratinho
```

Em certo tipo de documentos a função que a pessoa preenche no documento permite também ao tradutor gerar automaticamente informação relacional, além da informação funcional. Nos baptismos e casamentos os tradutores utilizam a informação funcional registada pelo utilizador para gerarem as relações de parentesco implícitas entre os vários intervenientes. A mesma técnica aplicada ao exemplo acima produz um registo ainda mais económico:

```
testemunha$jose machado/m/obs=ja' apareceu em devassas anteriores.
      atr$ec/s
      atr$residencia/soure%vila
      atr$idade/24
      pai$manuel fernandes
            atr$alcunha/ratinho
```

Neste exemplo a relação de parentesco entre José Machado e Manuel Fernandes é automaticamente gerada durante o processo de tradução, assim como as relações funcionais entre ambas as pessoas e o acto em que aparecem. Note-se que o sexo de Manuel Fernandes se infere de modo automático.

O fragmento acima gera as linhas das tabelas da base de dados referidas no capítulo anterior:

Quadro 3.28 Linhas na tabela de pessoas (repetição)

Pessoas

id	*nome*	*sexo*	*obs*
d1692-t1	jose machado	m	
d1692-r1	manuel fernandes	m	

Quadro 3.29 Linhas na tabela de atributos (repetição)

Atributos

id	*pessoa*	*atributo*	*valor*	*data*	*fonte*	*obs*
d1692-t1c1	d1692-t1			16921004	d1692	

Quadro 3.30 Linhas na tabela de relações (repetição)

Relações

id	*origem*	*destino*	*tipo*	*valor*	*data*	*fonte*	*obs*
d1692-r1r1	d1692-r1	d1692-t1	parentesco	pai	16921004	ldp1692	
d1692-t1r1	d1692-t1	dev1692	funcao em acto	teste-munha	16921004	ldp1692	
d1692-r1r2	d1692-r1	dev1692	funcao em acto	referido	16921004	ldp1692	

Em todos as infomações registadas podemos utilizar os "aspectos" da linguagem kleio. Este exemplo inclui, na residência de José Machado, uma anotação desse tipo, utilizando-se o sinal "%" de percentagem para indicar a forma original da informação no documento. Podemos igualmente utilizar o sinal "#" para inserir um comentário arbitrário. Estas informações adicionais são lidas pelos tradutores e colocadas no campo "obs" da entidade a que estão associadas, como se vê nos resultados da tradução do exemplo anterior.

Os tradutores reconhecem ainda uma convenção importante sempre que se regista uma pessoa. Se, ao registarmos uma pessoa, tivermos a certeza que a pessoa em questão já existe na base de dados com determinada matrícula, podemos indicar essa informação ao tradutor. A vantagem de o fazermos é evitar que se gere mais uma linha na tabela das pessoas. Deste modo apenas são acrescentado os novos atributos e relações de uma pessoa existente, caso os haja. Por exemplo, José Machado tem a matrícula "d1692-t1" na nossa base. Se o utilizador fixou, por alguma razão, essa matrícula e mais tarde, ao registar outro documento, depara com o mesmo José Machado, pode indicar essa

informação ao tradutor usando um elemento especial na transcrição da pessoa, o elemento "mesmo_que".

```
pessoa$jose machado/mesmo_que=d1692-t1.
    atr$residencia/soure
```

A sequência "mesmo_que" indica aos tradutores que não é necessário acrescentar nova linha à tabela de pessoas, uma vez que a pessoa já se encontra registada. Neste caso o tradutor limitar-se-ia a gerar uma nova linha para a tabela de atributos guardando o par residencia/soure referido à pessoa "d1692-t1".

Ao utilizarmos o elemento "mesmo_que" efectuamos uma *identificação explícita.* Embora o grosso da identificação de pessoas seja feito só depois de todos os dados introduzidos, por vezes é útil registar no momento da transcrição de documentos que temos a certeza que duas pessoas são as mesmas. Essa certeza pode vir de informação que não é registável, por exemplo a caligrafia idêntica de duas assinaturas. O mecanismo "mesmo_que" permite registar essa certeza no momento da transcrição. As identificações explícitas têm contudo os seus problemas. No sistema actual o principal óbice reside no facto de que é difícil desfazer uma identificação explícita que mais tarde se tenha descoberto ser errada. É necessário separar os atributos e relações de uma pessoa erradamente criada em duas pessoas distintas. Assim a identificação explícita deve ser reservada só para casos muito seguros.

3.3.4.2.Fontes e actos

Normalmente um registo em Kleio é organizado como uma sucessão de actos transcritos num mesmo texto. Só actos bastante extensos, como as devassas, são registados isolados num único texto.

O caso mais típico é exemplificado pelos registos paroquiais. Cada texto da transcrição kleio inicia-se por um grupo que identifica a fonte.

```
baptismos$tipo=reg paroquiais/valor=baptismos 1689
             /localizacao=fol. 63-/ano=1689
```

Este grupo identifica a fonte original do ponto de vista arquivístico. Em Soure mantivemos os ficheiros de actos paroquiais organizados por anos. Uma média de 100 baptismos por ano fornecia uma transcrição de tamanho razoável. O mesmo para os óbitos. O número mais reduzido de casamentos é compensado pelo facto de a quantidade de informação por cada acto ser bastante maior. Para cada novo ano começavamos um novo texto com um cabeçalho como o mostrado acima.

Dentro de cada texto sucedem-se os actos. Um exemplo dos primeiros baptismos do ano de 1689:

```
b$b1689.1/2/1/1689/0/0/0/manuel cordeiro
n$helena/f
      pai$manuel francisco
            atr$residencia/paleao
      mae$isabel simoes
      pad$manuel gaspar
      mad$helena
            atr$ec/s
            pmad$paulo ribeiro


b$b1689.2/12/1/1689/0/0/0/manuel cordeiro
n$francisco/m
      pai$manuel simoes
            atr$residencia/casais dos cavaleiros
      mae$maria jorge
      pad$manuel domingues
            atr$residencia/eureca
      mad$isabel francisca
            mrmad$manuel fernandes
                  atr$residencia/cavaleiros
```

Cada baptismo inicia-se com o grupo "b$" a que se segue um número de ordem do baptismo, fornecido pelo utilizador. Utilizamos a convenção de preceder com um "b" o ano do baptismo e acrescentar, separado por um ponto, um número sequencial (b1689.1, b1689.2,....). A seguir ao baptismo segue-se a data do mesmo, e a data de nascimento da criança se disponível. No período de tempo que este trabalho enquadra o pároco rarissimamente registava a data de nascimento e daí a sequência de três zeros (dia,mês,ano) que vemos no cabeçalho dos baptismos. Finalmente vem o nome do celebrante.

Após o cabeçalho vêm as várias pessoas envolvidas no acto. Como dissemos acima as pessoas são registadas em grupos Kleio cujo nome denota a função da pessoa no acto. As funções definidas para os baptismos são: n (criança), pai, mae, ppai (pai do pai), mpai (mãe do pai), pmae (pai da mae), mmae (mãe da mãe), pad (padrinho), mad (madrinha), pmad (pai da madrinha), mrmad (marido da madrinha), ppad (pai do padrinho), referido e referida. Quase todas as funções geram, durante o processo de tradução, relações de parentesco automáticas. O sexo das pessoas é também derivado da função que ocupam no baptismo. A única função cuja designação é neutra ("n" para criança) necessita por isso do sexo explicitamente registado, como se vê nos exemplos acima.

O mecanismo de registo dos casamentos é semelhante.

```
casamentos$tipo=reg paroquiais/valor=casamentos 1689
      /localizacao=99/ano=1689

cas$c1689.1/10/1/1689/?/manuel rodrigues da carreira
      noivo$manuel simoes
            pnoivo$manuel simoes
                  atr$residencia/casais do pinheiro
                  atr$morto/antes
            mnoivo$maria francisca
      noiva$catarina rodrigues
            atr$ec/v
            pnoiva$manuel rodrigues
                  atr$morto/antes
                  atr$residencia/casal do benzedor
            mnoiva$isabel carvalho
```

```
                  atr$morta/antes
      test$gaspar rodrigues de carvalho
      test$jose da rocha
      test$antonio cordeiro
      test$manuel leao
      test$catarina madeira/sexo=f
            rel$parentesco/mulher/manuel rodrigues
marmelo/cmtest10011689
            referido$cmtest10011689/manuel rodrigues marmelo/m
```

Tal como os baptismos os casamentos iniciam-se com um número de identificação seguido da data do acto. De seguida vem o local onde a cerimónia decorreu (neste exemplo o local não foi determinado e vemos um ponto de interrogação). Finalmente o nome do celebrante. As pessoas seguem-se com a respectiva função. Os casamentos permitem simplificações similares aos baptismos sendo possível registar facilmente os pais e avós dos noivos (com as funções pnoivo, mnoivo, ppnoivo (pai do pai do noivo), etc... Permitem adicionalmente que se registem cônjuges anteriores com as formas mulher1 (primeira mulher), pmulher1 (pai da primeira mulher), e assim sucessivamente até quatro cônjuges anteriores para ambos os noivos, se fôr necessário. A função "test" designa as testemunhas do casamento. Normalmente são do sexo masculino pelo que os tradutores, na ausência de registo explícito do sexo da testemunha inferem um homem. Quando de facto temos uma mulher como testemunha de casamento, como neste exemplo, temos que registar o sexo explicitamente. Os tradutores não são suficientemente sofisticados para inferirem o sexo a partir do nome das pessoas - não porque isso implique grandes dificuldades do ponto de vista da programação mas porque exige que o tradutor mantenha uma tabela de nomes que não é muito prática de gerir.

Repare-se ainda, no casamento demonstrado, nos atributos "morto/antes" associados a algumas pessoas. Trata-se da convenção utilizada para registar a informação que, na data do acto, esses indivíduos já tinham falecido.

O tipo de acto mais complexo transcrito para este trabalho, e que serve de teste aos limites da liguagem, é a devassa. A riqueza da informação da devassa e o seu carácter atípico colocam desafios muito grandes à informatização.

Um exemplo de devassa:

```
devassa$dev1698/7/6/1698/visdor=manuel soares
gouveira/secretario=jeronimo pimentel
      /cota=III\/D,4,4,48
      /obs=em muito mau estado

testemunha$manuel gomes/m
      atr$ec/c
      atr$profissao/?#ilegivel
      atr$residencia/novos
      atr$idade/58
      atr$assina/b

      caso$c1698-1/amanc
            acusado$alexandre gomes
                  atr$ec/s
                  atr$freguesia/pombal?
                  atr$residencia/paleao
            acusada$francisca
                  atr$alcunha/serrana
                  atr$ec/c#marido ausente
                  atr$residencia/paleao
                  atr$naturalidade/santiago da guarda#segundo a
testemunha
                        francisco rodrigues. Isso explicaria a alcunha
"serrana"

      acusa$c1698-1/amanc/literal=ela pariu uma crianca ha um ano + ou
-

      caso$c1698-2/amanc
            acusada$maria da folha
                  /obs=teria ido para Condeixa-a-Nova por volta de
1698
                              ver devassa de 1698
                  atr$ec/s

            acusado$manuel neto
                  atr$profissao/pedreiro
                  atr$ec/c
                  atr$freguesia/ansiao#teria ido residir para ansiao
por
                                    volta de 1698

      acusa$c1698-2/amanc/literal=teriam ido viver para condeixa-a-
nova

testemunha$antonio ferreira/m
      atr$ec/c
      atr$profissao/vive de seu trabalho
```

```
atr$residencia/paleao
atr$idade/45
atr$assina/+

acusa$c1698-1/amanc
```

O primeiro grupo "devassa$" tem a informação sobre o acto e a fonte. O primeiro "dev1698" é a matrícula de identificação do acto. De seguida temos a data e a informação sobre o visitador e o secretário, a cota do livro e uma observação sobre o seu estado de conservação.

Segue-se o conteúdo da devassa propriamente dita. Tal como no documento original a transcrição organiza-se como uma série de testemunhos transcritos. Cada testemunha fornece o nome e um conjunto de elementos de identificação que tratamos aqui segundo as regras normais para o registo de informação pessoal (o atributo "assina" regista a forma como a testemunha assina o seu depoimento; "+" significa de cruz, "b" significa uma assinatura legível).

As testemunhas relatam casos. Quando um caso é referido pela primeira vez por uma testemunha é aberto um grupo "caso$". Nesse grupo vai se concentrar a informação que progressivamente se revelará sobre o caso e respectivos acusados.

Os casos têm um código de identificação único. Esse código servirá para registar acusações das testemunhas seguintes. Os casos têm igualmente um tipo (amancebamentos, usura, violência,etc...) para os quais usamos uma abreviatura.

Seguindo imediatamente o registo do código e tipo de caso vem a informação sobre os acusados. Mais uma vez utilizam-se os mecanismos normais de registo de pessoas.

Segue-se finalmente a acusação sob o grupo "acusa$" que regista a informação especifica desta testemunha sobre o caso. A diferença entre o registo de caso e o registo de acusação é a seguinte: o caso regista a informação

específica ao caso em si, que consiste na sua matrícula identificadora, o seu tipo e os acusados. Cada caso é registado uma única vez junto com a testemunha que primeiro o denuncia. Quando testemunhas posteriores acrescentam informação relevante sobre o caso completa-se o registo. Por exemplo pode acontecer que a primeira testemunha que denuncia um amancebamento identifique só parcialmente os intervenientes. Informação adicional fornecida pelas testemunhas posteriores é acrescentada ao registo inicial, normalmente com um pequeno comentário identificando a origem da informação.

A acusação regista os facto de determinada testemunha ter delatado determinado caso e utilizamos para esse efeito o grupo "acusa$". Este grupo limita-se a indicar o caso em questão junto com detalhes sobre a terminologia usada (elemento "literal") e qual origem do conhecimento que a testemunha tem do assunto (elemento "origem").

A transcrição de devassas representa, com alguns tipos de escrituras notariais, uma das formas mais complexas de utilização da linguagem Kleio. Contudo, pensamos que o resultado continua legível e a sua utilização em arquivo fácil. É um exemplo da flexibilidade da linguagem mas também do paradigma utilizado de conciliação entre um modelo de dados formalizado e a necessidade de uma transcrição próxima da fonte.

3.3.4.3.Conclusão

A construção dos tradutores permite ter o melhor de dois mundos. Uma metodologia de transcrição que produz textos muito legíveis e próximos da fonte original e uma estrutura muito formalizada de dados. O procedimento de tradução desses textos nos formalismos da base de dados relacional é feito automaticamente, evitando trabalho e, sobretudo, erros. O ponto fraco desta

aproximação é que os tradutores, na sua versão actual, não são programas genéricos. São especialmente construídos para determinado tipo de fonte. Não são programas fáceis de produzir apesar de extremamente eficazes uma vez feitos. O custo da sua produção para uma fonte nova varia com o grau de especificidade dessa fonte. Em geral um novo tradutor parte de um tradutor existente, modificado para atender às particularidades do novo tipo de documentos. De qualquer modo elaborar um novo tradutor para uma nova fonte é uma tarefa de programação relativamente especializada.

Contudo não pensamos que seja impossível elaborar tradutores genéricos. As versões actuais destes programas foram elaborados numa fase de desenvolvimento do sistema em que a meta-informação ainda não fazia parte da base de dados. Com a meta-informação formalizada e acessível podemos imaginar que descrições do formato dos documentos podiam ser incluídas na base de dados e assim criar um tradutor genérico que consultaria primeiro a base de dados e depois efectuaria a tradução. Este é um dos pontos do sistema que mais requer atenção no sentido de produzir uma arquitectura versátil e facilmentea adaptável a novas fontes. Actualmente a construção de um novo tradutor é uma tarefa morosa e especializada que exige um programador experiente.

3.4. CRUZAMENTO NOMINAL

3.4.1. INTRODUÇÃO

Designamos por cruzamento nominal a operação que agrupa as várias referências a uma mesma pessoa que se encontram dispersas em variadas fontes. Em regra o principal elemento que guia esse processo de agrupamento é o nome, daí o adjectivo que qualifica a operação[24]. Do ponto de vista informático, que nos ocupa aqui, este processo consiste em examinar a base de dados e criar um índice que forneça, para cada uma das pessoas reais, o

[24] A expressão em inglês é "nominal record linkage" que traduzimos por "cruzamento nominal". Embora, como é evidente, muitos atributos pessoais sejam utilizados no processo de identificação, o termo cruzamento nominal tem a vantagem de recordar a abordagem nominalista cuja apologia foi tão brilhantemente feita por Ginzburg e Poni: Ginzburg, C.; Poni, C., "Il nome e il come: scambio ineguale e mercato storiografico". *Quaderni Storici* ,Anno XIV ,Facicolo I, gennaio-aprile,pp. 181-190, 1979.

conjunto das informações disponíveis sobre essa pessoa dispersas por várias fontes.

A operação de consolidar biografias individuais a partir de informação dispersa é algo de extremamente comum na investigação histórica. O processo atinge contudo uma dimensão que pede reflexão metodológica quando se aplica a um conjunto significativo de pessoas. Nesses casos as questões de rentabilidade e fiabilidade tornam-se importantes. Um tipo específico de cruzamento nominal que desde cedo suscitou uma reflexão metodológica aprofundada foi a reconstituição de famílias feita pelos demógrafos. A dimensão do problema e a necessidade de produzir resultados tanto quanto possível comparáveis levou a uma normalização precoce, sob a forma de *métodos*, dos quais o método Henry-Fleury constitui a primeira e mais seguida incarnação[25].

Apesar da simplicidade da sua definição, a operação de cruzamento nominal é, na maioria dos casos, um processo extremamente complexo e moroso. Desde cedo a questão da sua automatização foi posta. O primeiro campo em que essa necessidade surgiu foi em Medicina. A necessidade de, por um lado, consolidar os registos informatizados de diferentes instituições da área da saúde e, por outro, o imperativo de recorrer a reconstituições das filiações genéticas de indivíduos com várias gerações de profundidade para o despiste

---

[25] Fleury, Michel; Henry, Louis, *Nouveau Manuel de dépouillement et d'explotation de l'état civil ancien,* 3ème édition, Paris, Ed. Institut National d'Études Démographiques, 1985. Esta terceira edição incorpora já uma reflexão sobre a utilização de computadores na reconstituição de famílias, embora a nível puramente hipotético. A exposição do método foi clarificada e dividida para "*permettre aux informaticiens d'établir leurs programmes d'après les perfomances des ordinateurs*", p.10. Esta preocupação tinha surgido nos anos 70: *Simulation d'une reconstitution de familles par ordinateur*. "Annales Demographie Historique", (1972), p. 303-309 e no mesmo número *Variations des noms de famille et changements de prènon. Problèmes qui en résultent pour le couplage automatique des données.*, p.245-250.

de determinadas doenças, levou os especialistas a utilizar os meios informáticos ou mecanográficos que se encontravam à sua disposição com o fim de cruzar grande número de registos. Este movimento começa logo nos anos 60 com a vulgarização dos primeiros computadores em instituições médicas[26].

Logo desde o início o problemarevelou se longe de ser trivial. Embora o ponto de partida fossem populações contemporâneas, com traços de identificação mais precisos do que os disponíveis para os indivíduos do passado, a escassez da informação registada, os erros frequentes de registo ou declaração, as variações nominais que são importantes em determinados contextos[27], cedo convenceram os pioneiros que o processo estava longe de ser simples.

Os historiadores não demoraram muito tempo a reconhecer esta área como crucial e capitalizar o pioneirismo da medicina: em 1973 é publicada uma obra que ainda hoje constitui uma referência central: *Identifying people in the past*,

---

[26] Segundo Schofield, em *Identifying People in the Past*, a melhor entrada no tema era Acheson, E.D. (ed), *Record linking in Medicine*, Oxford, Oxford University Press, 1968, obra que não nos foi possível consultar. As primeiras formalizações, ainda hoje utilizadas, nomeadamente no que diz respeito à quantificação das similitudes de atributos estão igualmente associadas a aplicações médicas: Newcombe, H. B. et al. - *Automatic linkage of Vital Records*.. "Science", 130 , (October 1959), p.954-959; Newcombe, H. B., *Record Linking: The Design of Efficient Systems for linking Records into Individual and Family Histories.* "American Journal of Human Genetics",19, 3, part 1 (1967), p. 335-359. As aplicações médicas do cruzamento nominal automatizado vulgarizaram-se ao ponto de existir pelo menos um programa especificamente comercializado para esse fim. Ver para uma perspectiva recente: Jaro, Matthew A. - *Probabilistic linkage of large public health data files.* "Statistics in Medicine", vol.14 (1985), p. 491-498.

[27] Nas populações norte-americanas a forte componente migratória cria uma deriva ortográfica e grandes variações nominais provocadas por "americanização" de nomes de origens diversas.

editada por E.A.Wrigley, onde se reunem os primeiros exemplos de aplicação dessas metodologias a dados históricos[28]. Comparando com as aplicações na área da saúde as maiores diferenças residiam nos tratamentos associados à exiguidade da informação que o tempo deixou ao historiador e a necessidade de ultrapassar variações ortográficas muito mais extremas.

Os grandes emprendimentos historiográficos de cruzamento nominal datam dessa altura e estão intimamente associados às grandes máquinas e às ferramentas de desenvolvimento de software disponíveis nos anos 70. É o caso da reconstituição automática de famílias pelo Cambridge Group que servirá a English Population History[29] e do sistema SOREP do Centre Interuniversitaire de Recherches sur les Populations, no Canada[30].

Nesse tempo os custos de implementar um sistema automatizado de cruzamento nominal, ou mais frequentemente, de reconstituição de famílias, era extremamente elevado. Os orçamentos dos projectos tinham de enquadrar a acessoria especializada necessária ao desenvolvimento de software num contexto onde as ferramentas de programação e as soluções pré-fabricadas eram limitadíssimas[31]. Em consequência destes constrangimentos o

---

28 Wrigley, E.A. (ed), *Identifying People in the Past*, London. Edward Arnold, 1973.

29 Wrigley, E.A.; Schofield, R.S - *English Population History from Family Reconstitution: Summary Results 1600-1799*.. "Population Studies", 37 (1983), p. 157-184

30 Bouchard, Gérard; Roy, Raymond; Casgrain, Bernard - *Reconstitution automatique des familles: le systéme SOREP*. Chicoutimi: Université du Québec, 1985. e mais recentemente Bouchard, Gérard, *Current issues and new prospects for computerized record linkage in the province of Québec*. "Historical Methods" vol. 25, nº 2 (Spring 1992), p. 67-73.

31 Winchester, Ian,*What every Historian needs to know about record linkage for the microcomputer era*.. "Historical Methods", vol.25 (Sep 1992), p.149-. Nesse contexto o aparecimento do primeiro e primitivo software de gestão de bases de dados em micro-computadores, nos finais dos anos 70, como o célebre DBASEII, constituiu uma revolução. Antes disso todas as manipulações de base de dados que não fossem triviais necessitavam de programadores altamente especializados. O sistema canadiano SOREP era programado em Fortran, uma

desenvolvimento de soluções tendeu a concentrar-se na forma mais eficaz de resolver os problemas específicos a cada projecto e raramente na produção de procedimentos genéricos. Por regra é mais difícil e dispendioso produzir um sistema geral do que um sistema desenhado para resolver uma situação concreta. Sistemas que tenham ambições totalizantes, visando uma grande variedade de fontes com o objectivo genérico de cruzar "toda" a informação disponível, são particularmente dispendiosos (em termos de tempo e meios necessários) de implementar.

Não podemos esquecer que, no tipo de projecto a que nos estamos a referir, o objectivo principal era sempre de tipo historiográfico e não metodológico. Partes cruciais da automatização, como a representação da informação, o tratamento e resolução das variações ortográficas, os mecanismos de aferimento e resolução das ambiguidades, foram resolvidos num contexto de máxima eficiência local sem preocupação de generalidade. Com um estatuto instrumental dentro de projectos historiográficos cujos fins eram essencialmente outros, os programas de computadores para o cruzamento nominal e respectivos procedimentos acessórios nunca foram produzidos com um intuito de transportabilidade.

Temos conhecimento de dois projectos de ambições genéricas : o sistema Kleio desenvolvido pelo Max-Plack-Institut für Geschichte em Göttingen e o projecto de informatização da reconstituição de Earls Colne, dirigido por Alan MacFarlane. Tratam-se de projectos de natureza diversa. O sistema Kleio é um programa de gestão de base de dados especificamente desenhado para historiadores e que pretende ser uma ferramenta genérica. O projecto de MacFarlane pretendia informatizar toda os dados referentes a uma paróquia

---

linguagem mais apropriada ao cálculo matemático do que à gestão de bases de dados flexíveis. As ferramentas e as máquinas determinam antes do mais uma certa "filosofia" do sistema que configura o cenário de utilização, o tipo de utilizador e as extensões futuras.

inglesa que tinha sido manualmente reconstituída[32]. Ambos tiveram que enfrentar a questão da generalidade das fontes. Enquanto que o sistema Kleio se tem desenvolvido e é utilizado em várias instituições espalhadas por diferentes países o projecto de MacFarlane foi abandonado sem ter produzido ferramentas reutilizáveis. Tanto quanto sabemos em ambos os casos a questão do cruzamento nominal não foi resolvida satisfatoriamente. O sistema Kleio inclui dispositivos de tratamento de variações ortográficas que auxiliam a identificação de pessoas e a criação daquilo que aqui denominámos um dicionário de pessoas reais. Mas trata-se de na prática de uma identificação "manual" auxiliada por computador. De qualquer modo é natural que a evolução do sistema Kleio se faça no sentido de incorporar capacidades cada vez mais sofisticadas de cruzamento nominal.

A divulgação do sistema Kleio é um exemplo das consequências da evolução tecnológica para a generalização de ferramentas de trabalho apropriadas aos historiadores. A partir dos finais dos anos setenta a divulgação dos micro-computadores tendeu a democratizar um poder de cálculo que antes estava confinado a grandes instalações de acesso difícil. À primeira vista este facto deveria ter levado à popularização de programas que suportassem o cruzamento nominal, sobretudo na sua concretização mais comum, a reconstituição de famílias.

---

[32] A melhor apresentação do projecto de MacFarlane dentro da problemática que nos interessa aqui é feita por Robert Rowland em *L'informatica e il mestiere dello storico*. "Quaderni storici". Vol. 78 (1991), p. 693-720. O projecto Kleio tem produzido documentação própria. Consultámos Manfred Thaller, *Kleio: a Database System,* Max-Planck-Institut für Geschichte, St. Katharinen, 1993. Este manual não refere operações relacionadas com cruzamento nominal. Completámos com *A Tutorial for Kleio*, documentação facultada durante uma acção de formação que teve lugar no Instituto Universitário Europeu de Florença em 1993, a que tivémos acesso graças à gentileza de José Pedro Paiva.

Mas para que uma verdadeira generalização das ferramentas ocorresse seria necessário um investimento significativo na generalização dos métodos e esse investimento não ocorreu. Apesar da tecnologia ter evoluído no sentido de tornar cada vez mais comum o tipo de poder de cálculo que os pioneiros tiveram à sua disposição, outros elementos essenciais de um programa genérico de cruzamento nominal dependiam da vulgarização de ferramentas de software que só recentemente se tornaram verdadeiramente acessíveis. É o caso dos sistemas de bases de dados relacionais potentes, capazes de gerir eficazmente dezenas ou centenas de milhares de registos, que hoje podemos executar em computadores pessoais mas que há apenas quatro anos exigiam grandes sistemas e um nível de especialização humana muito grande

As bases de dados relacionais vão permitir solucionar os problemas de organizar e tratar a grande quantidade de informação que está sempre subjacente aos empreendimentos de cruzamento nominal. Com a difusão de bases de dados relacionais baseadas numa linguagem de interrogação e manipulação de dados padronizada, divulgaram-se também novas ferramentas de desenvolvimento de programas capazes de lidar com grandes quantidades de informação[33]. Estas evoluções do software foram tão decisivas como o puro aumento de poder de cálculo das máquinas. O mesmo se pode dizer das metodologias de construção de programas que procedem "inteligentemente" combinando informação de valor variável. A tecnologia dos sistemas periciais, que desenvolveu soluções para problemas que envolvem decisão sobre informação incerta e complexa, é igualmente bastante recente, ou pelo menos

[33] Sobre bases de dados relacionais ver acima 3.2.2.2.

posterior aos grandes projectos referidos anteriormente (à excepção do sistema Kleio).[34]

Assim as verdadeiras condições técnicas que permitem o desenvolvimento de programas genéricos para o cruzamento nominal começam lentamente a desenhar-se no fim da década de 80, tornando-se a partir daí cada vez mais potentes e fáceis de usar. Este processo faz baixar o nível de perícia necessária ao desenvolvimento de um sistema de cruzamento nominal genérico[35].

O objectivo deste capítulo é apresentar procedimentos que levam ao cruzamento nominal de informação sobre pessoas registada na estrutura de dados apresentada no capítulo 3.2. As soluções aqui propostas foram pensadas tentando maximizar a generalidade das aplicações. Ao contrário de outros

---

[34] Sobre sistemas periciais ver:Forsyth, Richard (ed), *Expert Systems. Principles and Case Studies*, London, Chapman & Hall, 1984. A literatura sobre o assunto é contudo vastíssima. Particularmente úteis no contexto desta investigação foram as referências especializadas em sistemas periciais aplicados a grande quantidades de dados e nos aspectos probabilísticos do "raciocínio inteligente" automático: Kerschberg, Larry (ed)-*Expert Database Systems, Proceedings from the First International Workshop*. (Benjamin/Cummings Series in Database Systems and Applications). Menlo Park, Ca, 1986 e Neapolitan, Richard, *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Serie: (Willey-InterScience Publication). New York, John Wiley & Sons Inc., 1989.

[35] Um primeira tentativa de programa genérico de cruzamento nominal que funciona sobre uma base de dados comercial (DBASE ou compatível) é o Matchmaker: Atack, Jeremy; Bateman, Fred; Gregson, Mary E. - *Matchmaker, Matchmaker, Make Me a Match: A General Personal Computer-Based Matching Program for Historical Research*. "Historical Methods", 25, 2, Spring 1992. O software está disponível gratuitamente por ftp em ftp.vanderbilt.edu/pub/pc. A empresa Matchware technologies vende software de características genéricas para cruzamento nominal que é utilizado em projectos da área da saúde, inclusivamente no Brasil. Ver informação disponível em http://www.matchware.com. O sistema Kleio, já referido, inclui igualmente mecanismos de auxílio ao cruzamento nominal embora, na versão sobre a qual temos informação, estes se limitem ao tratamento de variações ortográficas por algoritmos do tipo SOUNDEX.

programas, estes foram desenhados desde o início com o intuito de poderem servir noutras circunstâncias, noutras fontes. Nem sempre foi possível garantir um sistema completamente geral, mas pensamos ter avançado significativamente nessa direcção.

A generalidade dos algoritmos de cruzamento nominal que iremos apresentar está intimamente ligada à generalidade da estrutura de dados apresentada anteriormente. Essa estrutura fornece uma base versátil para os procedimentos adicionais necessários para identificar pessoas. Implica também, algumas limitações que identificaremos ao seu devido tempo. Como o cruzamento nominal vai produzir nova informação, um índice de todas as ocorrências de cada pessoa na documentação, iremos ter necessidade de estender a estrutura de dados já apresentada com novas estruturas. Iremos também mostrar como se lida com a informação temporária produzida pelo processo de identificação.

O resto da secção aborda vários aspectos ligados à construção de um sistema automatizado de cruzamento nominal e está estruturada do modo que explicitamos de seguida.

Em primeiro lugar forneceremos uma visão geral do mecanismo de cruzamento nominal aproveitando para explicitar os conceitos básicos do processo e definir um vocabulário preciso para descrição das várias entidades e fases envolvidas.

O mecanismo de identificação de pessoas tem uma base probabilística importante. O recurso às probabilidades e à teoria da informação fornece um suporte formal a muito daquilo que constitui a intuição do operador humano. Os programas necessitam de ter uma representação de quais os atributos mais raros e de quais os mais frequentes, quer sejam nomes, profissões, locais de residências ou cargos oficiais. Com base na raridade de cada atributo é possível

quantificar o efeito da concordância ou discordância do mesmo em duas pessoas que tentamos identificar. Por outro lado sabemos igualmente que determinados atributos são mais voláteis que outros, isto é, assumem vários valores ao longo da vida de um indivíduo. Podemos por isso valorizar diferentemente o facto de esses atributos concordarem ou discordarem. Estas operações implicam uma análise prévia cujo fim é elaborar as tabelas de frequência necessárias à estimativa das probabilidades associadas a cada valor de cada atributo. Exigem também uma análise *a posteriori* das pessoas já identificadas para determinar o grau de variabilidade real dos atributos. Descreveremos o modo como estas análises podem ser efectuadas sobre as estruturas de dados já descritas.

O cruzamento nominal utiliza intensivamente a informação registada sobre as pessoas. Devido ao elevado grau de automatismo que pretendemos atingir, a utilização que será feita da informação disponível é tão completa. Assim será necessário garantir que a informação constante da base de dados foi verificada, na medida do possível. É também necessário analisar os valores dos vários atributos pessoais no sentido de detectar eventuais variações ortográficas ou a existência de formas alternativas de designação. Embora os mecanismos de cruzamento nominal possam lidar com este tipo de variações, a qualidade da preparação prévia dos dados é determinante no resultado final do processo identificação. Assim serão descritos procedimentos prévios de normalização e os programas informáticos que os apoiam.

Finalmente o processo de identificação propriamente dito repousa sobre um mecanismo clássico dos algoritmos de inteligência artifical denominado "geração e teste". Os candidatos a serem uma mesma pessoa são recolhidos na base de dados (fase de geração) e os seus atributos analisados no sentido de determinar se todos são de facto ocorrências da mesma pessoa ou não (fase de teste). Este processo torna-se bastante complexo devido ao número de

comparações que se têm de fazer e também devido ao método a utilizar para decidir o que fazer quando encontramos candidatos incompatívéis. . Este é um dos problemas mais difíceis do cruzamento nominal. A metodologia seguida no nosso caso será explicada com a esperança que possa generalizada e melhorada noutras situações.

Finalmente veremos como foram implementados mecanismos de aferição da qualidade das decisões tomadas automaticamente e como esses mecanismos exigem a intervenção do historiador para os casos mais problemáticos encontrados durante o cruzamento automático.

De seguida iremos ver uma descrição genérica do processo de cruzamento nominal ao mesmo tempo que introduziremos o vocabulário conceptual básico indispensável para podermos mais detalhadamente examinar as soluções aqui definidas.

3.4.2. CONCEITOS BÁSICOS E FASES PRINCIPAIS DO CRUZAMENTO AUTOMÁTICO DE REGISTOS.

A complexidade de alguns dos processos ligados ao cruzamento nominal exigem precisões terminológicas indispensáveis para aclarar o significado dos conceitos base.

Designamos por "ocorrência" de uma pessoa o conjunto de informações sobre alguém que ocorre numa fonte. Numa base de dados temos várias "ocorrências" que corresponderão à mesma pessoa real. O conceito de "pessoa real" designa precisamente as pessoas que de facto existiram e deixaram vários traços (ocorrências) em diversas fontes. O objectivo do cruzamento nominal é de agregar as "ocorrências" em "pessoas reais". Mais concretamente, a

identificação de pessoas produz um índice que, a cada pessoal real, faz corresponder todas as "ocorrências" dispersas. A esse índice chamamos "dicionário de pessoas reais".

Uma ocorrência de pessoa que está a ser considerada para efeito de identificação designa-se por "candidato". Cada processo de identificação de uma pessoa real é lançado a partir de uma ocorrência inicial. A partir dos atributos dessa ocorrência é efectuada uma pesquisa de candidatos adicionais. Cada candidato é comparado com a ocorrência de partida e calculada uma medida do grau de conformidade dos respectivos atributos. Quanto mais atributos em comum existirem mais alta será essa medida e vice-versa. Se o valor da conformidade dos atributos ultrapassar um limiar pré-definido o candidato é aceite, senão é rejeitado.

Quando o candidato é aceite temos uma "ligação" entre a nova ocorrência e a ocorrência inicial. Essa ligação tem um valor associado: a medida do grau de conformidade dos atributos respectivos. Designamos esse valor ou medida por "valor da ligação". A determinação correcta do valor da ligação é essencial em qualquer sistema de identificação automatizado. Quando em fase posterior do processo fôr necessário tomar decisões sobre quais os candidatos a reter definitivamente e quais não, a análise do "valor" das ligações será central.

Partindo do candidato original irão ser recolhidas deste modo um certo número de ocorrências, cada uma delas associada a um "valor". Como a pesquisa foi feita a partir dos atributos da pessoa original é possivel que os candidatos recolhidos possuam informações adicionais que permitam relançar a pesquisa e obter ainda mais referências. Assim o mesmo processo é refeito para cada um dos novos candidatos, na esperança de obter mais ocorrências da pessoa em questão. Este processo designa-se por "fase de geração" e o seu resultado é um conjunto de candidatos inter-ligados por "ligações" e respectivos "valores".

Designamos por "regras de geração" as várias estratégias seguidas para recolher candidatos. Uma regra de geração normalmente baseia-se no nome da pessoa para encontrar todas as pessoas com o mesmo nome. Se o nome fôr demasiadamente comum então teremos outra regra que combina o nome com outro atributo adicional, por exemplo a residência. Se existirem outros atributos que auxiliem a identificação e forem suficientemente raros pesquisa-se a base para recolher todas as pessoas com esses atributos e retêm-se como candidatos aqueles que pelo menos tenham o primeiro nome em comum com a ocorrência de partida. Assim contornamos sempre que possível o efeito negativo que poderia ter uma variação nominal no apelido de uma pessoa, que é muito frequente em mulheres.

As regras de geração incluem mecanismos de filtragem. Assim podem nunca gerar candidatos separados por mais de um número de anos. Podem filtrar candidatos que produziriam ligações com um peso abaixo de um limiar pré-determinado. É boa metodologia de implementação fazer tantos testes quanto fôr possível na fase de geração, uma vez que o processamento necessário para lidar com eventuais candidatos espúrios posteriormente é sempre pesado. Assim procura-se optimizar cedo a eliminação de candidatos dentro do limite do razoável.

A afinação destas estratégias tem um papel central na economia geral da automatização do cruzamento nominal. De facto as estratégias de geração têm de corresponder a duas solicitações contraditórias: por um lado assegurar que não escapa nenhum candidato significativo; por outro impedir que se agreguem demasiados candidatos espúrios. O primeiro caso corresponde ao fenómeno de "sobre geração" e o segundo ao de "sub-geração". O equilíbrio é extremamente difícil de conseguir. O que é perfeitamente claro é o seguinte: não há nenhum modo eficaz de restringir o cruzamento nominal à simples afinação do processo de pesquisa e filtragem de candidatos. É intrinseco à

natureza do problema que o processo de geração produza candidatos a mais que têm de ser posteriormente separados em diferentes pessoas reais. Esta última fase só é possível ser implementada com eficácia se recolhermos todas as ocorrências que *eventualmente* possam corresponder à mesma pessoa.

Muitos dos verdadeiros testes à validade de cada candidato só podem ser feitos depois de todas as ocorrências recolhidas. Com efeito, grande parte do conhecimento que permite decidir, perante um conjunto de candidatos, se estamos perante uma só pessoa ou não, depende menos da comparação de pares de candidatos e mais da apreciação global do conjunto dos mesmos. Por exemplo, é necessário examinar a evolução do atributo estado civil no conjunto dos candidatos para sabermos se são todos compatíveis. Sequências do tipo solteiro-casado-viúvo-casado são aceitáveis enquanto que solteiro-casado-solteiro-viúvo já não são. Outro exemplo: podemos ter filtrado os candidatos garantindo que cada ocorrência só era considerada se tivesse menos de 80 anos de diferença da ocorrência inicial, mas por este processo podemos recolher um candidato que ocorreu 79 anos antes e outro 79 anos depois, somando 158 anos de diferença. Finalmente, aspectos como os intervalos intergenésicos devem ser considerados para detectar sobre-geração.

O conjunto de candidatos obtido pela fase de geração é normalmente tratado como uma rede, ou mais precisamente um grafo. Um grafo é um objecto matemático composto por uma série de elementos (nós) ligados por relações (arestas). Um exemplo frequentemente dado é a rede ferroviária. Trata-se de um grafo em que os nós são as estações e as arestas os troços de linha. O facto dos grafos terem uma representação computacional bem conhecida torna a sua utilização frequente em aplicações informáticas[36]. No

[36] O conceito de grafo e a sua importância no contexto dos processos que aqui se apresentam serão explícitados com mais detalhe nas páginas seguintes.

nosso caso os conjuntos dos candidatos a uma pessoa real constituem um grafo, em que cada um dos nós é uma ocorrência de pessoa e cada uma das arestas uma ligação de identificação com um peso determinado. Assim designaremos esses conjuntos de candidatos por "grafos de ligação". Muito do trabalho de identificação de pessoas é realizado sobre estes grafos. Os grafos de ligação constituem a representação formal de base para a informação com que se lida na identificação de pessoas.

Em conclusão é sempre necessária uma fase de testes sobre o grafo dos candidatos que passaram pela regras de geração e respectivos filtros. Essa fase baseia-se num segundo tipo de regras, as "regras de teste". Se o grafo de ligações passar todos estes testes então temos em princípio uma pessoa identificada. Se de facto se vier a demonstrar que ligámos a mais ou a menos, então será necessário rever as regras de geração e de teste. Mas o problema é realmente complicado quando as regras de teste detectam que o grafo não é coerente e contém ocorrências de pessoas que não podem corresponder à mesma pessoas real. Por outras palavras, as regras de geração juntaram várias pessoas no mesmo grafo. Isso significa que temos de dividir os candidatos em dois ou mais grafos que sejam, cada um, coerentes e passem todas as regras de teste. A complexidade do problema liga-se ao elevado número de soluções alternativas que são possíveis. O modo de o fazer não é simples e tem sido alvo de alargada discussão na literatura. Designa-se esta fase final do processo de identificação automática de pessoas, que só ocorre quando os grafos de ligação incluem mais que uma pessoa real, de "fase de resolução de ambiguidades".

Um resumo dos principais termos agora introduzidos ajudará a seguir o que se segue:

Resumo dos termos utilizados

candidato: ocorrência que está a ser considerada para inclusão numa pessoa real.
dicionário de pessoas reais: índice que fornece para cada pessoa real a lista de todas as ocorrências detectadas.
grafo de ligações: conjunto de todas as ligações entre todas as ocorrências que num dado momento estão a ser consideradas como candidatas à mesma pessoa real; resulta da pesquisa de todas as pessoas com atributos semelhantes.
ligação: relação entre duas ocorrências que possivelmente correspondem à mesma pessoa real. Ver valor
ocorrência:referência a uma pessoa na documentação
pessoa real: pessoa identificada por agregação de todas as ocorrências a ela referentes.
regras de geração: regras que a partir de um candidato inicial vão pesquisar ocorrências com atributos semelhantes; correspondem a instruções de procura na base de dados.
regras de teste: regras que aplicadas ao grafo das ligações detectam a existência de ocorrências incompatíveis, como por exemplo, diferença de mais de 100 anos entre baptismo e morte.
resolução de ambiguidades: procedimento pelo qual um grafo que não passou as regras de teste é segmentado em sub-grafos correctos.
valor da ligação: valor numérico que mede a plausibilidade de duas ocorrências pertencerem de facto a uma mesma pessoa real; valor calculado com base no número de atributos semelhantes, na raridade dos mesmos e no modo como variam ou não em situações reais.

Nas secções seguintes iremos aprofundar e detalhar estes conceitos e os procedimentos a ele associados.

3.4.3. Estratégias para um cruzamento genérico

Apesar de o cruzamento nominal automático ser um tópico recorrente e difundido com literatura especializada própria, quando descemos ao detalhe das implementações compreendemos que sob essa designação genérica se incluem situações diversas que implicam problemas e soluções diferentes. O facto de aqui se pretender abordar o problema de uma forma genérica coloca questões que outro tipo de projectos não tiveram de enfrentar.

Uma das dificuldades centrais que a generalidade do nosso objectivo coloca diz respeito à estratégia de cruzamento, isto é, o modo como os programas irão pesquisar a base de dados no sentido de agrupar todas as ocorrências em pessoas reais. Existem outras dificuldades, como é evidente, mas grande parte

delas decorre do problema de representar de forma geral qualquer tipo de fonte e qualquer tipo de informação biográfica. Essa questão já foi tratada em 3.2 e as soluções aí propostas constituem um capital importante para a resolução das questões que agora se nos vão levantar.

A questão das estratégias de cruzamento tem a ver com outro tipo de problema. Existem características de determinadas fontes que podem determinar uma estratégia de geração e teste optimizada. Por exemplo, suponhamos que a nossa base de dados incluíria apenas dois róis de confessados, de anos sucessivos. Pelo conhecimento que temos dos róis de confessados sabemos que cada pessoa só aparece num dado rol uma única vez. Assim o cruzamento nominal desta base de dados é muito simplificado. Basta comparar cada pessoa de uma das listas com um conjunto de candidatos da outra lista e escolher aquele que tiver uma ligação mais forte. Não temos obviamente que comparar as pessoas dentro de cada rol para saber se ocorreram mais do que uma vez. Tão pouco teremos que ter dúvidas sobre quantas pessoas num rol corresponderão a uma pessoa noutro. Sabemos que é uma correspondência de um para um. Finalmente sabemos também que neste tipo de populações é altamente provável que a maior parte das pessoas de um rol se encontre no outro, tirando as mortes, os nascimentos e as migrações. Podemos por isso ter uma expectativa informada do resultado do cruzamento.

Num sistema genérico estas expectativas e simplificações são mais difíceis. No que diz respeito aos registos paroquiais enquanto fontes temos alguns constragimentos importantes. Sabemos que as pessoas só se baptizam uma vez e só morrem uma vez. Mas podem aparecer como pais em baptismos um número não determinável de vezes (embora sujeito a constrangimentos). Podem igualmente casar-se mais do que uma vez. Quando estendemos o tipo da informação aos livros notariais e às visitas pastorais e ao mesmo tempo registamos e tratamos padrinhos e testemunhas nos baptismos e casamentos

chegamos a uma situação em que é impossível elaborar uma definição prévia de uma estratégia de cruzamento baseada em constrangimentos simples do tipo do que evocámos com o exemplo dos dois róis de confessados.

No contexto das reconstituições automáticas de famílias a questão é resolvida adoptando uma abordagem centrada nos actos. Por outras palavras o que se compara são actos e a comparação das pessoas é subsidiária da comparação de actos. Os vários projectos que recorreram a métodos informáticos para a reconstituição de famílias delinearam uma estratégia própria de elaboração do cruzamento. Por exemplo os algoritmos do Grupo de Cambridge sequenciam as comparações entre actos de um modo cuidadosamente planeado[37].

A sequenciação de comparações, limitando a procura de candidatos a ficheiros específicos, só é possível em situações em que o tipo de actos ou de fontes tem características muito particulares. É o caso das situações em que existem apenas registos paroquiais utilizados no contexto da reconstituição de famílias ou ainda do cruzamento de róis paroquiais ou de listas de censos.

No nosso caso, como foi referido, não existem constrangimentos muito claros. Não só o tipo de fonte é variado, como a utilização dos registos paroquiais inclui o tratamento de padrinhos e testemunhas de casamentos. No

---

[37] Ver Wrigley, E.A.; Schofield, R.S, *Nominal record linkage by computer and the logic of family reconstitution* . In: Wrigley, E.A. (ed), Identifying People in the Past. London. Edward Arnold, 1973, p 69-70. A sequência começa com a comparação entre os baptismos e o casamento dos pais, depois entre casamentos e óbitos com menção de casal, e assim sucessivamente. A ordem das comparações permite minimizar os testes e optimizar os processos. Permite também lidar com alguns problemas característicos dos dados ingleses como sejam o facto de as mulheres mudarem o apelido quando casam o que significa que não vale a pena comparar baptismos a óbitos sem ver primeiro as ligações baptismos/casamentos e casamentos/óbitos.

caso dos padrinhos e madrinhas de baptismo a quantidade de menções de casais introduzidas na base de dados é significativa.

Assim, não é rentável fazer uma abordagem centrada nos actos com sequências rígidas de comparações. Embora se pudesse fazer a reconstituição das famílias antes de uma identificação mais geral essa decisão corresponderia a uma perca considerável de informação. O facto de termos uma estrutura genérica de dados, como foi descrita na secção 3.2, implica uma estratégia global para o cruzamento.

A estratégia a seguir aqui só pode ser centrada nas referências individuais. O sistema é desenhado para encontrar todas as referências a uma pessoa partindo de uma ocorrência qualquer. Partindo de uma ocorrência inicial o sistema pesquisa todas as pessoas como o mesmo nome ou com um nome semelhante e para cada hipótese encontrada calcula a probabilidade de se tratar da mesma pessoa. Se essa probabilidade ultrapassar um limiar pré-definido então recolhemos o candidato para futuro exame. A maior parte das vezes, para identificar uma pessoa concreta, o sistema tem de considerar várias possibilidades que levam, na prática, à identificação de várias pessoas simultaneamente. Contudo o processo é claramente centrado na ocorrência singular e desenhado para, a partir de uma referência qualquer, encontrar todas as outras ocorrências pertinentes. O modo como o programa interage com o utilizador favorece também essa aproximação, permitindo que em qualquer momento possa ser lançado um processo de identificação sem que seja necessário processar toda a base de dados e cruzar todas as referências. É aliás perfeitamente possível trabalhar com dados parcialmente cruzados e refazer a identificação sempre e quando fôr necessário.

3.4.4. ASPECTOS PROBABILÍSTICOS E ESTATÍSTICOS

3.4.4.1. Da possibilidade de quantificar a semelhança de duas ocorrências

A automatização do cruzamento nominal depende fortemente da capacidade de representar numericamente a "confiança" que duas ocorrências de pessoas na documentação digam realmente respeito à mesma pessoa real. Certamente que, à primeira vista, a possibilidade de efectuar esse tipo de formalização parece remota. Os historiadores familiarizados com o cruzamento de fontes ou reconstituição de comunidades sabem que as decisões são muitas vezes tomadas numa base mais "intuitiva" do que exacta.

A questão de formalizar ou quantificar uma actividade inteligente humana é central nas aplicações de inteligência artificial, em particular na construção de sistemas periciais que pretendem reproduzir o raciocínio de um perito. O problema nem sempre é posto em termos exclusivamente técnicos. Frequentemente, desloca-se para um campo epistemológico onde as interrogações giram em torno do que é formalizável e do que não é. Esta é aliás uma discussão antiga e recorrente à volta da teoria das probabilidades[38].

Certamente que a muitos a ideia de que podemos quantificar a certeza de que duas referências em fontes dizerem respeito à mesma pessoa parecerá excessiva ou condenada ao fracasso. Podemos assegurar que aqui só se utilizou o grau de formalização mínimo necessário para resolver os problemas que empiricamente se foram levantando. É nossa convicção que muito do que se considera a intuição do historiador no processo de identificação de pessoas é o

---

[38] Neapolitan, Richard, *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. 1989, em especial o capítulo 2 "Probabilistic considerations" onde se compara as aproximações "subjectivistas" e "frequencistas" da teoria das probabilidades e onde se discute, de forma bastante clara, o papel da teoria das probabilidades na modelização do raciocínio humano.

resultado de uma absorção inconsciente das características essenciais da informação sob análise e que esse processo pode ser reproduzido em computador. Assim acreditamos que as certezas por detrás da maior parte das comparações são quantificáveis e pensamos poder demonstrá-lo empiricamente.

De seguida iremos explicitar os conceitos probabilísticos sobre os quais se constrói o método de quantificação da semelhança. O que iremos apresentar assenta nos princípios base criados desde os anos sessenta pelos pioneiros do cruzamento nominal automatizado. Contudo aqui esses princípios serão integrados num formalismo genérico de representação de informação que foi explicitado acima, na secção 3.2 e explicados não só em termos teóricos mas também nas vertentes que se prendem com a sua aplicação prática.

3.4.4.2. Quantidade de informação de um atributo

A perícia do historiador a identificar pessoas em várias fontes não é fruto apenas de capacidades intelectuais específicas do género humano. É seguramente o resultado de uma prática continuada de trabalho com as fontes. O historiador acabado de entrar em contacto com um conjunto documental relativo a uma comunidade que não conhecia, dificilmente toma decisões sobre quem é quem, a não ser em casos extremamente claros e inequívocos. Por outro lado, o investigador que durante anos se debruçou sobre uma grande variedade de fontes relativa a uma população, acaba por ser capaz de "intuições" esclarecidas. Em Soure, por exemplo, durante o processo de transcrição dos registos paroquiais, era possível adivinhar quem eram os padrinhos de determinados baptismos tendo apenas conhecimento do nome da criança. Não existe nada de transcendente nestes processos. São o resultado de

uma contabilização inconsciente da frequência com que ocorrem determinados atributos presentes na informação. É pelo lento labor de recolha de dados e transcrição de documentos que o historiador se torna um verdadeiro especialista na identificação de pessoas de determinada comunidade.[39]

Sabemos isso intuitivamente. A certeza que temos que dois nomes digam respeito à mesma pessoa real dependente estreitamente da concepção interior que temos da sua raridade. Duvidarei que dois "António Rodrigues de Soure" sejam a mesma pessoa mas não duvidarei que "Paulo Ribeiro Cabral, da Quinta de S.Tomé" seja sempre a mesma pessoa em documentos diferentes. Assim damos um valor maior à informação mais rara e um valor menor à informação mais comum.

Este procedimento intuitivo liga-se a princípios conhecidos da teoria da informação e da teoria das probabilidades. Em termos da teoria da informação esta situação pode ser descrito do seguinte modo: informação é tudo aquilo que diminui a nossa incerteza sobre determinado acontecimento. Quando uma pessoa é referida numa fonte o acontecimento sobre o qual a incerteza se refere é o seguinte: qual a pessoa real que está a ser referida? Supondo que a população em observação é de cerca de 10.000 pessoas, a nossa incerteza sobre o acontecimento é inicialmente de um para 10.000.

A informação necessária para eliminar esta incerteza é quantificável em termos da teoria da informação: é igual ao logaritmo base 2 do número de desenlaces possíveis para o acontecimento e expressa-se na unidade "bits".

---

[39] Assim teríamos dificuldade em aconselhar uma relação mediatizada do historiador com as fontes primárias no contexto de uma reconstituição de comunidades. O contacto directo com os registos paroquiais, pelo menos, para alguns anos de registo, é fundamental para absorver o "vocabulário" e construir inconscientemente uma hierarquia de pertinência da informação nominal, toponímica e outra.

Sendo o número de desenlaces 10.000 a quantidade de informação necessária a uma certeza absoluta está entre 13 e 14 bits[40].

O nome da pessoa é uma informação que reduz a nossa incerteza sobre quem está a ser referido num documento mas não a elimina completamente. Para a eliminar seria necessário que o nome identificasse univocamente cada pessoa. Seriam necessários assim 10.000 nomes diferentes, o que obviamente não é o caso. Saber o nome de uma pessoa reduz a nossa incerteza, como é evidente, mas quanto?

Obviamente que a resposta depende do quantidade de nomes diferentes existentes na população. Nos dias de hoje, em que nomes compostos de 5 partículas são frequentes, poderíamos esperar bastante poder descriminatório nos nomes. No passado, contudo, o espectro nominal é muito mais reduzido. Nomes de duas partículas são quase sempre a regra e um número reduzido de formas monopoliza grande parte dos nomes. Em Soure ocorrem 260 primeiros nomes diferentes e 660 apelidos. No total podemos contabilizar cerca de 2900 nomes completos diferentes para todas as ocorrências na base de dados.

Se tivermos cerca de 3000 nomes diferentes utilizados por uma população de 10000 pessoas podemos estimar que, em média, um nome é utilizado por cerca de três ou quatro pessoas. Isto seria verdade se os vários nomes se distribuissem uniformemente. Contudo, como sabemos, existem nomes mais comuns que outros. O quadro 3.31 mostra o número de vezes que ocorrem os nomes mais frequentes na base de dados.

---

[40] Um "bit" é a quantidade de informação que reduz a incerteza em 50%. Assim se um acontecimento tem dois desenlaces possíveis um "bit" de informação é suficiente para anular a incerteza. Se outro acontecimento tem 4 desenlaces possíveis são necessários 2 bits: um reduz a incerteza de 4 para 2 e o outro de dois para 1. É fácil de ver que a relação entre a quantidade de informação (i) e o total de desenlaces que determina a incerteza (n) é de $2^i=N$. Para dez mil desenlaces teríamos entre 13 e 14 uma vez que 2^{13}=8 192 e 2^{14}=16 384. Mais exactamente: 13,2977.

Quadro 3.31 Dez nomes mais frequentes (total de 33.000 referências nominais)

maria	1278
manuel	795
manuel rodrigues	587
isabel	494
maria rodrigues	458
jose	442
joao	362
joao rodrigues	345
maria francisca	344
manuel goncalves	333

A contagem do número de ocorrências de cada nome na totalidade da documentação levanta o problema de sobrevalorizar os nomes de pessoas que aparecem muito frequentemente na documentação, como sejam aquelas que detêm cargos oficiais ou que, pela sua importância dentro da comunidade, são frequentemente chamadas a servir de testemunhas ou apadrinhar crianças. Utilizemos por isso a lista dos nomes dos defuntos nos assentos de óbito, que não está sujeita a esta perturbação uma vez que em princípio não contém re-ocorrência de pessoas.

Quadro 3.32 Nos óbitos 1675-1720 (total 2754)[41]

maria	143
manuel	126
isabel	76
jose	59
antonio	55
joao	45
maria francisca	41
maria rodrigues	35
maria joao	32
antonia	29

[41] As diferenças entre as duas listas, no que diz respeito à ordenação relativa das várias formas não se deve ao fenómeno da re-ocorrência que não é relevante no grupo dos nomes mais frequentes. É antes precisamente nos nomes mais raros que os valores associados às pessoas mais visíveis se inflacionam.

maria simoes	29
isabel rodrigues	27
maria goncalves	27
domingos	26
isabel francisca	26
manuel rodrigues	24
.....	
josefa	16
.....	
anselmo	1

O facto de determinados nomes serem mais frequentes que outros tem consequências quantificáveis do ponto de vista da informação que transmitem. Suponhamos que temos a informação que alguém morreu e foi sepultado na paróquia de Soure entre 1675 e 1720, período para o qual possuímos 2754 óbitos registados. A nossa incerteza sobre qual a pessoa que morreu é de 1 para 2754. A quantidade de informação que necessito é, à partida,

$$i = \log_2(2754) = 11{,}4$$

Se a informação que possuímos nos diz que a pessoa falecida se chamava "Maria Francisca" então a nossa incerteza passa a 1 para 41 uma vez que encontro esse número de óbitos de pessoas com esse nome. Passámos de 2754 hipóteses a 41 hipóteses pelo facto de termos recebido a informação que a pessoa se chamava "Maria Francisca". Para diminuir a nossa incerteza totalmente necessitamos de mais informação que resolva essa dúvida residual,

$$j = \log_2(41) = 5{,}4$$

Posso assim concluir que a quantidade de informação veiculada pelo nome "Maria Francisca" é a diferença entre a incerteza **antes** de saber o nome e a incerteza **depois** de saber o nome:

$$i - j = 11,4 - 5,4 = 6$$

assim o nome "Maria" fornece 6 bits de informação.

Vejamos o que aconteceria se o nome do defunto fosse "Anselmo", que ocorre uma única vez nos óbitos daquele período. Como só temos um caso, a nossa incerteza é completamente reduzida ao conhecermos o nome. Assim a quantidade de informação adicional necessária **depois** de sabermos que o nome é "Anselmo" fica, evidentemente, reduzida a zero,

$$j = \log_2(1) = 0$$

Concluimos então que a quantidade de informação veiculada pelo nome "Anselmo" neste contexto é de:

$$i - j = 11,4 - 0 = 11,4$$

e que a quantidade de informação contida no nome "Anselmo" é maior que a que está associada ao nome "Maria Francisca".

Nomes diferentes veiculam quantidades de informação diferentes. Quanto mais raro é o nome maior a quantidade de informação que lhe está associada. Generalizando: quanto mais raros os atributos de uma pessoa mais informação transmitem. O princípio em si é evidente mas veremos que é extremamente útil a base formal que permite a sua quantificação.

Podemos expressar este princípio em termos de teoria das probabilidades[42]. Quanto mais provável fôr que um atributo ocorra por acaso na população, menor é a quantidade de informação que esse atributo transmite.

A probabilidade de que um óbito qualquer diga respeito a alguém chamado "Maria Francisca" é equivalente à proporção (frequência relativa) do nome dentro do total dos óbitos[43]:

$$P = 41/2754 = 0.0149 = 1{,}49\%$$

Do mesmo modo a probabilidade de que um óbito seja relativo a alguém chamado "Anselmo" é:

$$P = 1/2754 = 0.00036 = 0{,}036\%$$

---

[42] A obra de Neapolitan, já citada, fornece uma introdução formal à teoria da probabilidade. Útil pelo estilo simples e didáctico, que não supõe uma grande formação matemática anterior é: Winkler, Robert L,*An Introduction to Bayesean Inference and Decision*. (Seires in Quantitative Methods for Decision Making). New York, Holt Rinehart and Wiston, 1972, com uma introdução geral no capítulo 2.

[43] As probabilidades de um acontecimento ter determinado desfecho D exprimem-se por valores entre 0 (D não ocorre de certeza) e 1 (D ocorre de certeza). Se tirarmos um óbito ao acaso do conjunto de óbitos temos uma probabilidade 0,0149 que o nome do defundo seja "Maria Rodrigues" o que significa que repetindo o acontecimento muitas vezes poderíamos esperar encontrar esse nome 1,49% das vezes. Aceita-se em teoria das probabilidades que a frequência relativa da ocorrência de um acontecimento sirva como aproximação da probabilidade do acontecimento ocorrer, uma vez que ambos os valores tendem a coincidir quanto o número de tentativas é elevado — é a chamada "lei dos grandes números". Para maior legibilidade optámos por expressar a probabilidade em termos de percentagens sempre que apropriado. Diremos assim uma probabilidade de 50% em vez de 0,5.

Relacionando estes valores com os anteriormente encontrados chegamos a uma lei fundamental da teoria da informação:

a informação de um signo S é igual ao logaritmo da sua improbabilidade

ou seja:

$$\text{Inf(S)} = \log_2 \left(\frac{1}{\text{P}}\right)$$

ou, escrito de outro modo

$$\text{Inf(S)} = -\log_2 (\text{P})$$

em que P é a probabilidade de o signo (atributo) S ocorrer por acaso, que calculamos como a frequência relativa, ou seja, o número de vezes que o signo ocorre dividido pelo total de ocorrências.

Aplicando aos nossos exemplos:

Para "Maria Rodrigues"

$$\text{Inf(S)} = \log_2\left(\frac{1}{\frac{41}{2754}}\right) = \log_2\left(\frac{1}{0.0149}\right) = 6$$

Para "Anselmo"

$$\text{Inf(S)} = \log_2\left(\frac{1}{\frac{1}{2754}}\right) = \log_2\left(\frac{1}{0.00036}\right) = 11{,}4$$

Estes são os valores a que anteriormente tínhamos chegado formalizando a situação em termos de incerteza e informação. Existe assim uma ligação estreita entre os conceitos de incerteza, informação, acaso e probabilidades.

Atingimos assim uma primeira noção que é fundamental para a construção de um sistema automático de cruzamento nominal:

> A confiança no resultado da comparação de duas ocorrências é directamente proporcional à quantidade de informação contida nos atributos comparados, quantidade essa que, por sua vez, é proporcional à improbabilidade ou raridade dos mesmos.

3.4.4.3. Contabilização da variabilidade de um atributo: o príncipio dos pesos binários

A utilização do conceito de quantidade de informação não esgota, contudo, a problemática da quantificação da semelhança. A maioria dos sistemas de automatização de cruzamento nominal baseiam-se numa medida derivada da quantidade de informação denominada em inglês "binit weights", ou pesos binários. Os pesos binários são um modo de medir a quantidade de informação de cada atributo para efeitos de cruzamento nominal. São definidos do seguinte modo[44]:

> O peso a dar à comparação de um atributo comum para efeito de identificação é igual à razão entre a probabilidade de esse atributo ter o mesmo valor em ocorrências diferentes da mesma pessoa real e a probabilidade de o atributo ter o mesmo valor em ocorrências de pessoas diferentes.

---

[44] Para uma visão geral das várias medidas de similitude utilizadas em projectos de cruzamento nominal ver: Winchester, Ian - *A brief survey of the algoritmic, mathematical and philosophical literature relevant to historical record linkage*. In: Wrigley, E.A. (ed)-Identifying People in the Past.. London. Edward Arnold: 1973.

Este valor exprime-se como um logaritmo base 2, tal como a quantidade de informação.

$$\text{peso binário} = \log_2\left(\frac{P(A)}{P(B)}\right)$$

em que

P(A) probabilidade do atributo ter o mesmo valor em ocorrências da mesma pessoa

P(B) probabilidade do atributo ter o mesmo valor em ocorrências de pessoas diferentes.

Embora esta medida possa parecer menos evidente que o conceito de quantidade de informação anteriormente apresentado a sua necessidade explica-se facilmente. A raridade que está por detrás da quantidade de informação não determina por si só o valor de identificação de um atributo. Um aspecto adicional a ter em conta é a variabilidade. Há características das pessoas que são relativamente estáveis ao longo do tempo enquanto que outras podem variar com frequência. Parte destas variações devem-se a deficiências de registo, erros de transcrição e outras perturbações na forma como a informação chega até nós. Contudo em muitos casos trata-se de variações reais[45]. Por exemplo o nome é muito mais estável do que a residência. Certas profissões são mais estáveis do que outras. Um lugar de residência pode ser relativamente raro, por ser pouco habitado, mas ser ocupado por pessoas que

[45] Que a variabilidade é intrínseca aos dados nem sempre parece ser evidente. Por vezes este aspecto é identificado, redutoramente, como a taxa de erro do registo. Um exemplo é Jaro, Matthew A.,*Probabilistic linkage of large public health data files*. "Statistics in Medicine", vol.14, pp.491-498, 1995.

circulam rapidamente para outros locais. Assim para além da raridade temos que ter em linha de conta a variabilidade de um atributo.

A variabilidade dos atributos tem origens diversas. Em primeiro lugar existe uma variabilidade real que decorre da própria realidade social: as pessoas mudam de residência, podem mudar de profissão e inclusivamente mudar de nome com relativa frequência. Em Soure as elites locais ligadas à governança possuem frequentemente duas residências, uma casa na vila e uma quinta nos arredores. Podem ser registadas indferentemente como residindo em Soure ou na quinta respectiva. Em segundo lugar diferentes fontes descrevem os atributos pessoais com graus diferentes de precisão. O povoamento extremamente disperso do território de Soure provoca uma fragmentação da toponímia que nem todas as fontes acompanham. Existe assim aquilo que designámos por "topónimos fracos", normalmente casais onde vivem uma ou duas famílias que, se estiverem na proximidade de uma aglomeração mais significativa, facilmente são "absorvidos", reaparecendo mais tarde numa fonte diferente. Assim a mesma pessoa, em fontes diferentes, pode aparecer como residindo nos "Caldeiros" ou nos "Simões" sem que de facto tenha havido mudança de residência. O que acontece é que dada a pouca importância do topónimo "Caldeiros" e à proximidade da aglomeração maior que é "Simões" o primeiro é frequentemente "absorvido" pelo segundo. Em terceiro lugar existe uma variabilidade que é fruto simultaneamente das mutações reais ocorridas e do modo como estas são registadas e tratadas posteriormente durante o processo de cruzamento nominal. Trata-se de atributos relacionados com relações de parentesco para as quais necessariamente haverá várias pessoas. Por exemplo o nome da mulher de um homem variará se esse homem recasar. Um caso ainda mais frequente é o nome dos filhos. Em fontes diferentes, como é óbvio, o mesmo pai ou a mesma mãe aparecem registados com filhos de nome diferente (só a série dos baptismos

fornece uma enorme quantidade de ocorrências desse género). Apesar do nome do filho não ser estritamente um atributo pessoal os programas de cruzamento automatico utilizam-no no processo de identificação, assim como, evidentemente, o nome dos conjuges. A probabilidade de uma pessoa real aparecer em duas ocorrências distintas como um filho de nome igual é muito baixa numa base de dados que inclua registos paroquiais, porque a esmagadora maioria das vezes que um pai ou mãe aparecem estão associados a filhos diferentes que sucessivamente baptizam, enterram ou casam. O facto de o nome dos filhos raramente ser o mesmo de ocorrência para ocorrência significa que não deve ser usado para efeitos de identificação de pessoas? Responder afirmativamente seria descartar uma informação que apesar das características apontadas tem por vezes uma importância significativa na resolução de certas ambiguidades. Poderíamos criar casos especiais em que o nomes dos filhos seria considerado, como o cruzamento entre os casamentos e o baptismo dos noivos mas isso seria criar um processamento especial que, como explicámos anteriormente, seria sempre uma solução parcelar no contexto de uma base de dados genérica como esta.

A solução está na fórmula do peso binário. Se um atributo fôr muito variável, como a residência no caso da população aqui analisada, então a probabilidade do atributo concordar em pessoas realmente identificadas diminui. O numerador da fórmula do peso binário desce e o valor geral da medida desce também. Inversamente atributos muito estáveis aumentam a probabilidade de o mesmo valor se encontrar em diferentes ocorrências da mesma pessoa e o peso binário aumenta. O nome dos filhos é um atributo de altíssima variabilidade e por isso a probabilidade de coincidir em duas ocorrências de uma mesma pessoa é muito baixa.

O princípio básico do cálculo do peso binário, acabado de expôr é, em conclusão, bastante simples. Trata-se afinal de um aperfeiçoamento do conceito de quantidade de informação inicialmente apresentado. Em determinadas circunstâncias, como veremos, os dois conceitos confundem-se e resultam no mesmo valor.

Se o conceito do peso binário é simples já a metodologia que leva à sua determinação é alvo de muita discussão[46]. O problema decorre da circularidade da definição: o peso binário serve para valorizar a comparação das ocorrências de pessoas para efeitos de cruzamento nominal mas exige que se conheçam valores que só estão disponíveis após o cruzamento nominal estar feito. Efectivamente, saber as probabilidades de um atributo variar ou não em diferentes ocorrências de pessoas reais exige um conjunto de pessoas identificadas previamente a partir do qual essas variações possam ser contabilizadas. Esse conjunto deve ser suficientemente significativo para que as probabilidades calculadas sejam fiáveis e úteis.

A solução para o círculo vicioso que daqui decorre varia bastante conforme os projectos. Uma das abordagens consiste em criar um conjunto de identificações manuais que possam servir de base à estimativa das probabilidades P(A) e P(B) da fórmula do peso binário. Mas em populações de tamanho considerável, onde o espectro de atributos seja significativo, essa tarefa pode revelar-se problemática e levanta todos os problemas que procedimentos de amostragem sempre colocam. Alguns programas simplesmente pedem ao utilizador uma estimativa *a priori* desses valores para

---

[46] Aliás a própria formulação do conceito de "binit weight" está sujeita a ligeiras mas frequentes variações na literatura sobre o assunto. Essas variações prendem-se, como veremos, à necessidade de proceder a simplificações e extrapolações no cálculo do valor real dessa medida. Ver Winchester, *op. cit.*.

cada um dos atributos e ensaiam o cruzamento automático a partir daí. Os resultados são revistos manualmente e validados ou recusados. Com base nessa revisão o sistema analisa as identificações consideradas correctas, corrige a estimativa das probabilidades de variação, e recomeça de novo. O processo pode ser repetido indefinidamente, criando, em princípio, cruzamentos progressivamente mais correctos[47] .

O problema de fazer amostragens significativas ou de validar grandes quantidades de cruzamentos reside na introdução de uma componente de decisão humana que alguns projectos consideram ou demasiado custoso ou perigosamente sujeita a erros e que, por isso, deve ser mantida em limites muito controlados. Muitos autores enveredam em consequência por abordagens que visam simplificar ou estimar os valores de P(A) e P(B) a partir da análise quase exclusiva das ocorrências de pessoas antes de qualquer processo de identificação. Com esse fim são feitas algumas simplificações e assumidas algumas equivalências apriorísticas. Estes métodos baseiam-se principalmente na análise das características matemáticas das probabilidades envolvidas e na importância relativa dos dois principais componentes do cálculo dos pesos binários: a raridade e a variabilidade dos atributos. Embora não seja possível ter um cálculo exacto da variação dos atributos sem um trabalho empírico que depende sempre de validação humana, é importante delimitar o papel desta intervenção. Trata-se de um ponto central na "respeitabilidade" dos processos de cruzamento automático e seguramente um dos aspectos que mais críticas suscita, porque, compreensivelmente, o efeito de

[47] É o método usado pelo programa "MatchMaker", já referido. Obviamente todos estes procedimentos que dependem quer de ficheiros cruzados manualmente quer de ensaios validados posteriormente dependem muito da qualidade do trabalho dos intervenientes humanos.

simplificações de natureza matemática sobre dados referentes a populações reais é sempre polémico.

O sistema aqui apresentado baseia-se de facto em algumas dessas simplificações. O processo de cruzamento nominal pode partir de um desconhecimento total prévio da variabilidade dos atributos e ir incorporando o conhecimento acumulado pelas identifiações sucessivas à medida que estas são validadas. Mas ao fazê-lo tenta isolar aquilo que depende de facto irremediavelmente de resultados validados pelo historiador daquilo que pode ser inferido de uma base de dados não ligada. A distinção entre os dois aspectos é subtil e longe de ser evidente. É aliás um dos pontos em que mais divergência provoca entre os vários projectos que utilizam a tecnologia dos pesos binários.

Para melhor compreensão dos raciocínios que levam à abordagem aqui seguida é necessário exemplificar o peso relativo da raridade e da variabilidade dos atributos e o modelo teórico que permite a simplificação desses valores a partir de dados reais. Para isso será necessário introduzir uma situação fictícia que permita criar um cenário suficientemente simples para demonstrar com clareza os príncipios em obra.

3.4.4.4. Interacção das várias componenentes quantitativas: frequências relativas, pesos binários e probabilidades condicionais

Nesta secção será explicado o modo como se chega ao processo de determinação do valor da concordância entre atributos de pessoas para efeitos de identificação. Como vimos esse valor é normalmente calculado como uma proporção entre duas probabilidades: a de o atributo concordar em ocorrências diferentes da mesma pessoa e a de concordar em ocorrências de pessoas diferentes. Como não possuímos à partida informação suficiente para

determinar esses valores temos de proceder por uma mistura de estimativas, amostragem e assumir algumas simplificações. Iremos descrever uma situação hipotética que tornará a interacção destes vários factores mais claros. O nosso objectivo é criar um modelo suficientemente simples para os factores principais serem compreensíveis.

Suponhamos, assim, uma população de 80 pessoas. Dessa população temos duas listas nominais obtidas em diferentes momentos. Para tornar a situação realmente simples vamos assumir que ambas as listas são exaustivas e não redundantes. Ou seja, em cada uma das listas aparece cada uma das 80 pessoas sem repetições. A função desta simplificação é sabermos claramente à partida qual o número de identificações correctas que se podem fazer.

Suponhamos ainda que nessa população de 80 pessoas só existem três nomes diferentes: os nomes A,B e C, que se distribuem do seguinte modo: o nome A é comum a 10 pessoas, o nome B a 20 e o nome C a 50 pessoas.

Numa situação normal teríamos obviamente um espectro de nomes mais variado e não teríamos tantas certezas sobre a dimensáo real da população nem sobre as características das duas listas nominais. A simplificação que fazemos aqui serve o propósito de tornar simples e claros os cálculos que se vão fazer.

O objectivo deste exercício é determinar o modo como calculamos o valor, para efeitos de identificação, de duas pessoas terem o mesmo nome.

Um sistema de cruzamento nominal terá que comparar os atributos das pessoas das duas listas para determinar quem é quem. Assim, em teoria, para cada pessoa de uma das listas irão ser comparadas todas as pessoas da outra lista. Mais uma vez, numa situação real, nuca se compararíam todas as ocorrências de pessoas entre si e tentar-se-ia limitar os pares a examinar. Aqui é

importante considerarmos o total de comparações teoricamente possíveis que são 80 vezes 80 ou seja 6400 pares de ocorrências.

Designemos uma das listas nominais por L e a outra por M. Designemos a primeira pessoa da lista L por L1, a segunda por L2 e assim sucessivamente até a última que será L80. Utilizemos o mesmo processo para designar as pessoas da lista M.

Ao compararmos todas as pessoas produzimos, como vimos, 6400 pares que podem ser representados pelos identificadores das pessoas que estão a ser comparadas. Temos assim os pares:

(L1,M1), (L1,M2), (L1,M3),..., (L1,M80),
(L2,M1), (L2, M2), (L2,M3), ..., (L2,M80),
.....
(L80,M1), (L80, M2), (L80,M3), ..., (L80,M80),
ao todo 6400 pares de ocorrências para as duas listas de 80 pessoas.

Destes 6400 pares só 80 correspondem a comparações verdadeiras, pois esse é o número de pessoas reais que existem e, sendo ambas as listas exaustivas, cada pessoa de uma delas aparece igualmente na outra. Imaginemos que as 80 pessoas ocorrem em cada lista pela mesma ordem, de modo que L1 é a mesma pessoa que M1, L2 a mesma que M2 e assim por diante. Deste modo os únicos pares que incluem duas ocorrências da mesma pessoa real são:

(L1,M1), (L2,M2), (L3,M3), ..., (L80,M80) [ao todo 80 pares]

Temos em conclusão 80 pares com ocorrências da mesma pessoa real e 6320 pares com ocorrências de pessoas diferentes. Para comodidade vamos designar os pares que dizem respeito à mesma pessoa real por "pares verdadeiros".

Podemos calcular a probabilidade de um par qualquer ser um par verdadeiro dividindo o número destes pelo total de pares:

Probabilidade de um par ser verdadeiro = 80/6400 = 1,25%.

Do mesmo modo:

Probabilidade de um par não ser verdadeiro = 6320/6400 = 98,75%.

Estes valores não são em regra conhecidos previamente em casos reais de cruzamento nominal. O nosso modelo apresenta claramente um valor alto para a probabilidade de um par ser verdadeiro, o que decorre da exiguidade da população imaginada. Por exemplo, o cruzamento de dois censos de uma população de 8000 pessoas poderia fornecer algo como 7000 pares verdadeiros e 57 milhões de pares referentes a pessoas diferentes. Como é claro, a probabilidade a priori de um par ser verdadeiro diminui com o número de ocorrências nominais envolvidas. Nesse caso seria de 7000/5.000.000. Inversamente, mas de modo muito menos significativo, essa probabilidade aumenta com o grau de recorrência de pessoas reais nas fontes utilizadas. Se poucas pessoas aparecem muitas vezes, a probabilidade de um par qualquer de corrências ser um par verdadeiro é maior.

Apesar de evidentes estas constatações preliminares são importantes para justificar as aproximações seguintes. Interessa reter, para já, que existe uma grande desproporção entre os pares de ocorrências que dizem respeito às mesmas pessoas reais e o conjunto de pares possíveis. Podemos representar esta proporção graficamente por um quadrado com uma área equivalente ao número total de pares e inserir o rectângulo correspondente ao número de pares verdadeiros (ver diagrama 3.4):

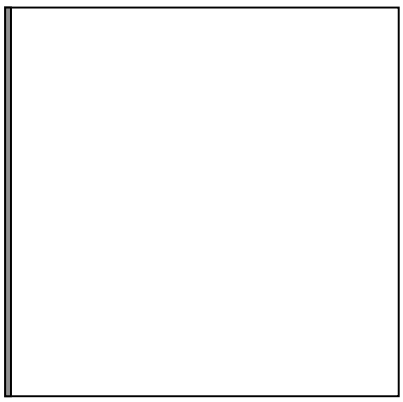


Diagrama 3.4. Probabilidade de um par de ocorrências ser verdadeiro em duas listas nominais completas e sem repetições para uma população de 80 pessoas. A zona sombreada à esquerda, quase imperceptível, representa a a probabilidade de 1.25%. Para um número não trivial de casos essa zona seria invisível a esta escala.

Que papel poderá ter o atributo nome na identificação das pessoas dentro deste modelo simplificado? Em primeiro lugar vejamos a distribuição dos vários nomes pelo conjunto de pares de ocorrências. O nome A ocorre em dez das pessoas reais. A sua frequência relativa dentro da população é de 10/80 ou seja 12,5%. O nome B tem uma frequência de 25%, ocorrendo 20 vezes e o nome C com 50 ocorrências constitui 62,5% dos casos.

Em termos de pares de ocorrências os cálculos são simples. Quando as dez pessoas com o nome A da lista L são comparadas com as 80 pessoas da lista M são gerados 800 pares que correspondem a 12.5% de todos os pares. Desses, 100 incluem ocorrências em que o nome A está duplamente presente. São os pares gerados pela comparação das dez pessoas com nome A em cada uma das listas. Os outros 700 incluem uma ocorrência de pessoa com o nome A e outra pessoa com o nome B ou C. Finalmente desses 100 pares em que o nome A aparece em ambas as ocorrências, apenas 10 correspondem a pares verdadeiros. Podemos elaborar uma tabela que relaciona o número de pessoas com cada nome com o

número de pares de diferentes características que se encontram no espaço total das comparações (ver tabela 3.33).

Tabela 3.33 número de pessoas com cada nome, frequência relativa e tipos de pares de ocorrências

Nome	Pessoas	Freq. Rel.	Pares	Nome =	Verdadeiros
A	10	12.5%	800	100	10
B	20	25%	1600	400	20
C	50	62.5%	4000	2500	50

Actualizando o diagrama 3.3 obtemos uma visualização clara das relações entre estes valores (ver diagrama 3.4).

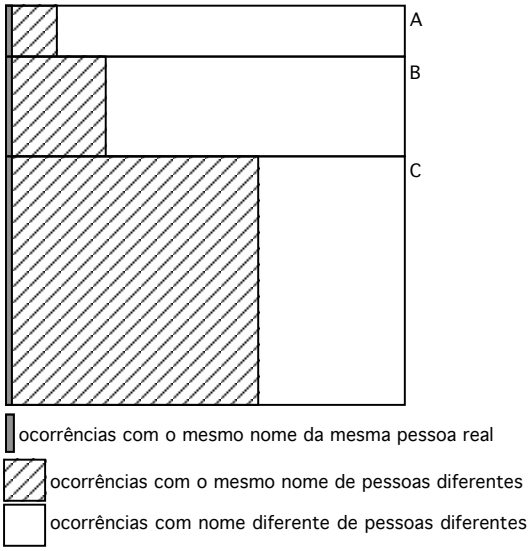


Gráfico 3.5 proporção dos vários tipos de ocorrências em comparações de duas listas de 80 pessoas com três nomes de frequência diversa. A razão entre a zona sombreada e a zona tracejada na diagonal aumenta à medida que o atributo se torna mais frequente e é de 1 para 10 em A e 1 para 50 em C.

A primeira constatação importante é que a proporção de pares verdadeiros, quer em relação ao total de pares gerados, quer em relação aqueles que unem

ocorrências com o mesmo nome, diminui à medida que a frequência do nome aumenta. Veja-se que a proporção de pares verdadeiros dentro dos pares com nome igual é de 1/10 em A, 1/20 em B e 1/50 em C. Isso significa que face a um par de ocorrências com o mesmo nome a probabilidade de estarmos perante a mesma pessoa é de 10% se o nome fôr A, 5% se o nome fôr B e 2% se o nome fôr C.

Esta probabilidade de uma pessoa ser a mesma dado o facto de o nome ser igual denomina-se uma probabilidade condicional, pois faz o desenlace de um acontecimento depender da ocorrência de outro. O acontecimento final, neste caso, consiste em um par de ocorrências dizer respeito a uma mesma pessoa. O acontecimento "propiciatório" consiste em, nesse par, as ocorrências terem o mesmo nome. O que pretendemos saber é: qual a probabilidade da pessoa ser a mesma dado o nome ser igual? Nesta simulação o cálculo é fácil porque temos uma ideia precisa do número de pares verdadeiros, o número de pares com o mesmo nome e o total de pares existentes.

O diagrama 3.5 permite-nos visualizar o processo pelo qual probabilidade condicional se calcula. Inicialmente a probabilidade de um par dizer respeito à mesma pessoa real consiste na chance do par em questão se situar na estreita banda sombreada do lado esquerdo do gráfico. Como vimos esse valor é igual a 80/6400 ou seja 1,25%. Contudo, se soubermos que o par tem o mesmo nome e que esse nome é A, então o nosso espectro de possibilidades reduz-se para o quadrado, no canto superior esquerdo, que representa o conjunto de pares em que A é atributo comum de ambas as ocorrências. Agora a probabilidade de o par ser verdadeiro consiste na chance de ele se situar na faixa sombreada dentro do quadrado referido. Esta nova probabilidade é de 10/100, ou seja 10%. O mesmo raciocínio aplicado aos nomes B e C produz as probabilidades já referidas de 5% e 2% respectivamente.

Do ponto de vista matemático uma probabilidade condicional deste tipo é calculada a partir da fórmula dada pelo Teorema de Bayes[48]:

$$P(V \mid N) = \frac{P(N \mid V) \bullet P(V)}{P(N)}$$

em que:

P(V | N) é a probabilidade de um par ser verdadeiro **dado** ter um determinado nome em ambas as ocorrências - o valor que procuramos.

P(N | V) é a probabilidade de que um par verdadeiro tenha esse nome em ambas as ocorrências.

P(V) é a probabilidade de um par qualquer ser verdadeiro independentemente dos nomes associados.

P(N) é a probabilidade de um par qualquer, verdadeiro ou não, ter esse mesmo nome em ambas as ocorrências.

A aplicação da fórmula ao caso do nome A torna mais claro o significado de cada um dos componentes.

P(V | N) é a probabilidade de um par ser verdadeiro **dado** o facto de ambas as ocorrências terem o nome A.

P(N | V) corresponde à probabilidade de um par verdadeiro ter o mesmo nome A em ambas as ocorrências. Como há 80 pares verdadeiros e 10 com o mesmo nome A a probabilidade é de 10/80=12,5%.

P(V), a probabilidade de um par qualquer ser verdadeiro, é 80/6400, ou seja, como já vimos, 1,25%.

---

[48] Winkler, Robert L. -*An Introduction to Bayesean Inference and Decision, (Seires in Quantitative Methods for Decision Making)*, New York: Holt Rinehart and Wiston, 1972. p.42.

P(N), a probabilidade de um par qualquer, verdadeiro ou não, ter o mesmo nome A é de $100/6400 = 1{,}56\%$.

Substituindo na fórmula temos:

$$P(V \mid N) = \frac{12.5\% \text{ X } 1.25\%}{1.56\%} = 10\%$$

valor a que tínhamos chegado empiricamente.

Por detrás da formulação matemática do cálculo da probabilidade condicional está um princípio relativamente simples e que se formula melhor com os conceitos de "hipótese" e "indício". A "hipótese" consiste na possibilidade de um par de ocorrências dizerem respeito à mesma pessoa. O "indício" consiste em o nome coincidir. O que nos diz a fórmula do teorema de Bayes? Em primeiro lugar parte-se da probabilidade da hipótese ser verdadeira antes de possuirmos qualquer indício — a chamada probabilidade *a priori*. É o valor P(V) acima e que corresponde a 1.25%. Seguidamente considera-se em que grau o índicio está presente quando a hipótese se confirma. Por outras palavras, qual a probabilidade de, dado que um par é verdadeiro, o nome coincidir? Designa-se esta probabilidade por $P(N \mid V)$. Multiplicamos estas duas probabilidades para obter o numerador. Assim o numerador $P(N \mid V)P(V)$ representa a probabilidade de a nossa hipótese se confirmar multiplicada pela probabilidade de o indício ser revelador da veracidade da hipótese. Ao dividirmos este produto pela probabilidade do indício ocorrer por acaso, P(N) , estamos a dimininuir a força do indício na probabilidade final de modo proporcional à sua vulgaridade. Em resumo, podemos formular assim a regra da probabilidade condicional expressa pelo Teorema de Bayes:

> A probabilidade de uma hipótese se confirmar dado determinado indício é proporcional à probabilidade da hipótese se confirmar por acaso e à probabilidade do índicio estar presente quando a hipótese se confirma, ao mesmo tempo que é inversamente proporcional a probabilidade do indício acontecer por acaso.

Repare-se que esta formulação é extremamente semelhante à definição do peso binário dada acima na secção 3.5.3. O peso binário, recordemos, é directamente proporcional à frequência com que o nome coincide em pares verdadeiramente ligados e inversamente proporcional à frequência com que o nome coincide em pares referentes a pessoas diferentes. Ora a "frequência com que o nome coincide em pares verdadeiramente ligados" é equivalente à "probabilidade do indício estar presente quando a hipótese se confirma", enquanto que "a probabilidade do indício ocorrer por acaso" é muito próxima da "frequência com que o nome coincide em pessoas diferentes" [49].

---

[49] Uma derivação da formula dos "binit weights" a partir do teorema de Bayes encontra-se em Hershberg, Theodore; Burstein, Alain; Dockhorn, Robert, *Record Linkage*, "Historical Methods Newletter", vol.9, nº2 & 3, pp.137-163, 1976. Outro autor, Winchester, ao rever os vários métodos de quantificar o contributo dos atributos semelhantes para a certeza de identificação, fornece a fórmula da probabilidade condicional transformada pelo logaritmo base dois (log2 (P(N,V)/P(N)) como equivalente ao cálculo dos pesos binários — Winchester, Jan, *The linkage of Historical Records by Man and Computer: Techniques and problems* , "Journal of Interdisciplinary History".,1 ,107-124, 1979. Ambas as publicações referêm-se a esta última fórmula como a ponte entre a inferência bayesiana tradicional e o cálculo dos "binit weights" e designam-a por "functor de confirmação de Hamlin" referindo uma dissertação inédita a que não tivemos acesso: Hamlin, C.L., *Language and the Theory of Information"*, dissertação de doutoramento apresentada à Universidade de Londres, 1955. De facto existem ligeiras diferenças conceptuais como veremos no texto mas os seus efeitos práticos são negligenciáveis.

Vejamos como o cálculo dos pesos binários se aplica nesta simulação.

Para o nome A:

P(A) = frequência da coincidência de A em pares verdadeiros = 10/80 =12,5%

P(B) = frequência da coincidência de A em pares falsos = 90/6320 = 1.42%

$PB_A = \log_2(P(A)/P(B)) = \log_2(12{,}5\%/1{,}42\%) =$ **3,134**

Para o nome B:

P(A) = frequência da coincidência de B em pares verdadeiros = 20/80 =25%

P(B) = frequência da coincidência de B em pares falsos = 380/6320 = 6,01%

$PB_B = \log_2(P(A)/P(B)) = \log_2(25\%/6{,}01\%) =$ **2,056**

Para o nome C:

P(A) = frequência da coincidência de C em pares verdadeiros = 50/80 =62,5%

P(B) = frequência da coincidência de C em pares falsos = 2450/6320 = 38,77%

$PB_C = \log_2(P(A)/P(B)) = \log_2(62{,}5\%/38{,}77\%) =$ **0,689**

Podemos visualizar as duas frequências ou probabilidades envolvidas no cálculo do peso binário utilizando o esquema já apresentado (ver diagrama 3.6).

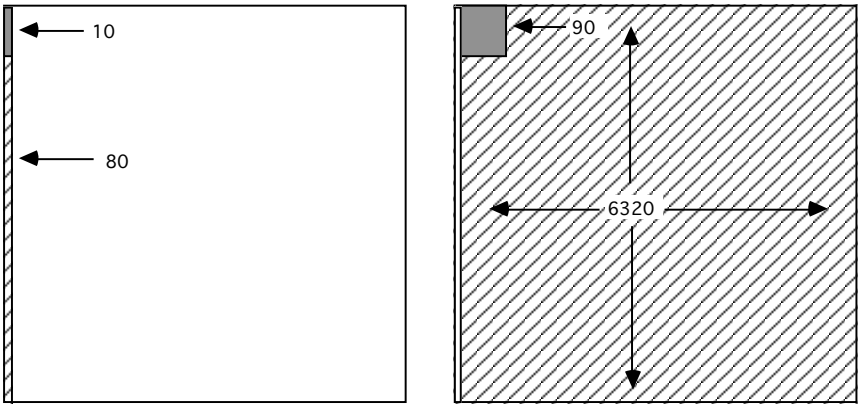


Diagrama 3.6 As duas probabilidades envolvidas num peso binário: à esquerda a frequência de ocorrência de pares com o nome A entre os pares verdadeiros; à direita a frequência de pares com o nome A entre os pares falsos

Com base na aplicação da fórmula do peso binário a este modelo, e mantendo presente a representação gráfica das proporções envolvidas chegamos finalmente ao ponto em que é possível introduzir as simplificações necessárias para a utilização destes formalismos em situações reais. Porque necessitamos de simplificar? Porque numa situação real nunca sabemos a priori a importância da faixa sobreada do gráfico 3.4, nunca sabemos à partida quais os pares verdadeiros. Sabemos apenas que o seu número, em relação ao total de pares é extremamente pequeno.

Introduziremos agora duas das simplificações base da utilização destes formalismos em situações reais de cruzamento nominal:

> Simplificação 1: a frequência da coincidência de um atributo em pares verdadeiros tende para a frequência relativa desse atributo na população em geral.

> Simplificação 2: a frequência da coincidência de um atributo em pares falsos tende para o quadrado da frequência relativa desse atributo.

Ambas as simplificações têm os seus riscos que explicitaremos mais adiante, mas são facilmente justificáveis.

Quanto à primeira simplificação o raciocínio é muito simples e fácil de visualizar no nosso modelo gráfico (ver diagrama 3.5). Se um nome, por exemplo A, ocorre em 12,5% das pessoas, ocorrerá na mesma proporção no total de comparações que são feitas — é essa a proporção da banda A que atravessa à largura a parte superior do gráfico. Por outras palavras: o total de pessoas com o nome A, que é dez, está para o total das pessoas, que é 80, na mesma proporção que o total de pares em que A aparece, que é 800, está para o total de pares possíveis, que é 6400:

$$10/80 = 800/6400 = 12{,}5\%$$

Como se vê no diagrama 3.5, a proporção mantém-se dentro da estreita faixa dos pares verdeiros, e mais importante, repare-se que a proporção seria a mesma se a largura da faixa fosse maior ou menor. O número de pares verdadeiros em que coincide o nome A é de 12.5% do total dos pares verdadeiros e esse valor seria o mesmo se o número de pessoas reais fosse maior ou menor (ver diagrama 3.7).

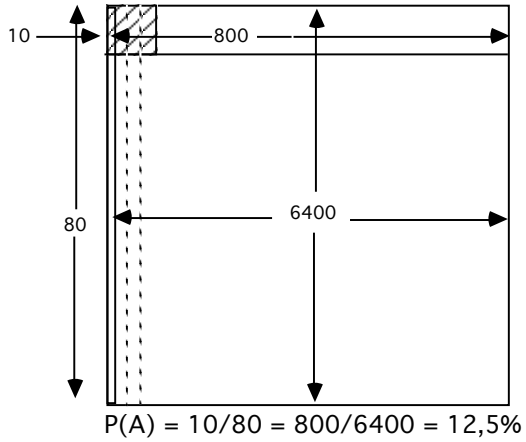


Diagrama 3.7: ilustração da simplificação 1: "a frequência da coincidência de um atributo em pares verdadeiros tende para a frequência relativa desse atributo na população em geral". A proporção de 10 para 80 é a mesma que 800 para 6400. Note-se que esta relação manter-se-ia mesmo que a proporção de pares verdadeiros no total do espaço de comparações fosse diferente (linhas verticais tracejadas).

Se a frequência dos atributos numa base por ligar puder ser considerada uma aproximação razoável da frequência dos atributos na população em geral então essa frequência pode ser utilizada como base da probabilidade do atributo coincidir em pessoas reais, fornecendo-nos assim o numerador da fórmula dos pesos binários. Esta asserção terá de ser relativizada pela consideração da possibilidade do atributo variar dentro de pessoas reais, fenómeno a que já aludimos anteriormente e que mais adiante examinaremos de novo no contexto deste modelo. Esta é uma das simplificação basilares dos procedimentos automáticos de cruzamento nominal.

Quanto à segunda simplificação, que a frequência de um atributo coincidir em pessoas que não são as mesmas tende para o quadrado da frequência relativa do mesmo, o príncipio subjacente prende-se com a desproporção entre o número de pares verdadeiros e o número de pares falsos. Veja-se gráficamente a diferença entre os dois valores (diagrama 3.8).

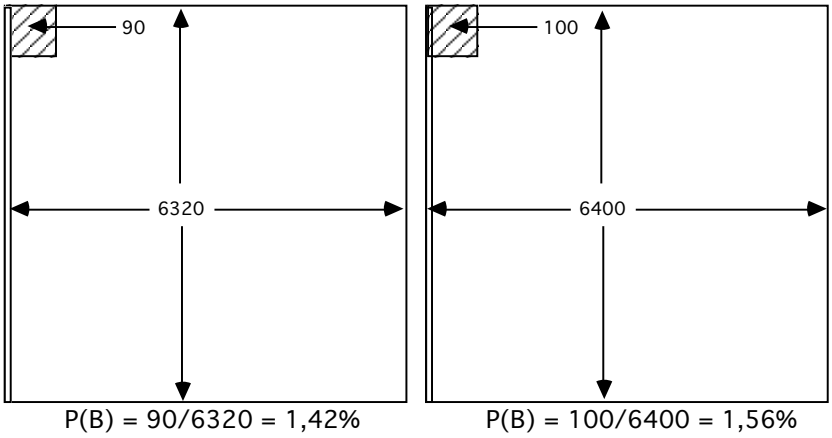


Gráfico 3.8: a pouca importância relativa do número de pares verdadeiros faz com que não exista grande diferença em usarmos o quadrado da frequência relativa sobre o total de pares (à direita) em vez da forma "correcta" que seria a proporção de pares falsos com o mesmo nome no total de pares (à esquerda).

A diferença entre os dois valores (1,42% e 1.56%) reside como se vê na consideração ou não dos pares verdadeiramente ligados. Contudo, como a proporção desses pares no total geral de pares é muito pequena, a diferença torna-se mínima. É, no nosso modelo, uma pequena diferença, mas numa situação real, em que a proporção de pares verdadeiros é muitissimo menor, a variação é negligenciável. Esta simplificação é obviamente muito importante porque apesar de termos uma ideia da ordem de grandeza da proporção de pares verdadeiros, não sabemos o seu número exacto a não ser no fim do cruzamento nominal. Sabemos contudo a frequência relativa de cada atributo pelo que podemos, por este modo, calcular o denominador da fórmula do peso binário[50].

Neste cenário, e com base nestas simplificações, poderíamos reescrever a fórmula do peso binário do seguinte modo:

---

[50] Esta simplificação é comum em outros projectos sendo por vezes apresentada como um facto evidente. Aqui quisemos dar uma justificação tanto quando possível clara da sua origem.

$$PB_A = \log_2\left(\frac{\text{frequência relativa de A}}{\text{frequência relativa de A}^2}\right)$$

Refazendo o cálculo podemos averiguar as diferenças finais das nossas simplificações sobre os resultados.

Para o nome A:

P(A) (estimada)= frequência relativa de A = 10/80 =12,5%

P(B) (estimada) = quadrado de P(A) = 12,5% ^{2} = 1.56%

PB_A(estimado)= $\log_2$(P(A)/P(B)) = $\log_2$(12,5%/1,56%) = **3,00**

PB_A(real)= **3,134**

Para o nome B:

P(A) (estimada)= frequência relativa de B = 20/80 =25%

P(B) (estimada) = quadrado de P(A) = 25% ^{2} = 6,25%

PB_B (estimado)= $\log_2$(P(A)/P(B)) = $\log_2$(25%/6,25%) = **2,00**

PB_B (real)=**2,056%**

Para o nome C:

P(A) (estimada)= frequência relativa de C = 50/80 =62,5%

P(B) (estimada) = quadrado de P(A) = 62,5% ^{2} = 39,06%

PB_C (estimado)= $\log_2$(P(A)/P(B)) = $\log_2$(62,5%/39,06%) = **0,689**

PB_C = $\log_2$(P(A)/P(B)) = $\log_2$(62,5%/38,77%) = **0,678**

Estes resultados permitem ajuizar a legitimidade das simplificações apresentadas. Os resultados finais são muito próximos e tenderiam a sê-lo mais se o número de casos fosse maior. Repare-se ainda que o cálculo do peso binário, tal como fica definido após as simplificações apresentadas, é

absolutamente equivalente ao conceito de quantidade de informação apresentada na secção 3.5.2. Com efeito, partindo de:

$$PB_A = \log_2\left(\frac{\text{frequência relativa de A}}{\text{frequência relativa de A}^2}\right)$$

e dividindo ambos os termos da fracção pela frequência relativa de A obtemos

$$PB_A = \log_2\left(\frac{1}{\text{frequência relativa de A}}\right)$$

ou seja, o logaritmo da improbabilidade de A — a sua quantidade de informação[51]. Assim, em determinadas condições, o peso binário e a quantidade de informação são a mesma coisa.

Esclarecer essas condições leva-nos ao elemento final que necessitamos de introduzir para cobrir a totalidade dos aspectos relacionados com o cálculo da semelhança: a **variabilidade** dos atributos dentro de diferentes ocorrências de pessoas reais. Essa variabilidade, como vimos na secção anterior, tem várias origens e é um factor importante a ter em conta quando se avalia o contributo de cada informação para a certeza final da identidade de diferentes ocorrências.

No nosso modelo, e no gráfico que o ilustra, estava implícito que dentro dos pares verdadeiros só existiam pares de nomes iguais, distribuídos pelos vários nomes. A estreita faixa dos pares verdadeiros sobrepunha-se aos quadrados a

---

[51] O mesmo raciocínio, embora sem referir o conceito de quantidade de informação, encontra-se em Newcombe, H. B., Kennedy, J.M.; Axford, S.J.; James, A.P.,*Automatic Linkage of Vital Records*."Science" ,130 , Outubro de 1959, pp.954-959.

tracejado que representavam os pares com o mesmo nome. Até agora raciocinámos apenas em função das proporções entre pares verdadeiros do mesmo nome e pares falsos com o mesmo nome. De certo modo, podemos dizer que temos os problemas derivados da *homonomia*, devidamente tratados. Falta-nos por isso tratar a *variação nominal*, ou seja a possibilidade de duas pessoas serem registadas com nomes diferentes sendo de facto a mesma pessoa. Este aspecto, recordamos, é generalizável a todos os atributos, sendo necessário admitir sempre que os valores podem discordar dentro de pessoas reais.

Para introduzirmos a variação nesta simulação temos que supôr que a faixa dos pares verdadeiros não cobre só os pares de nomes iguais mas também alguns pares de nome diferente. Podemos ilustrar o fenómeno graficamente (ver diagrama 3.9).

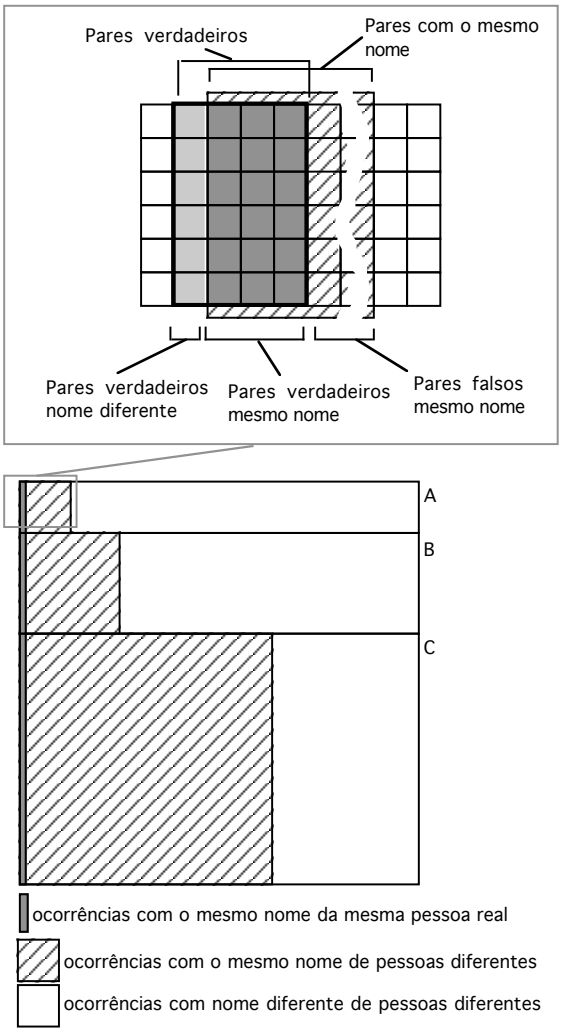


Diagrama 3.9: numa situação real os pares verdadeiros não contêm só casos em que o nome coincide. Em alguns casos a pessoa é a mesma apesar do nome variar. Esses casos são aqui representados a sombreado mais claro enquanto que os pares verdadeiros de nome igual são representados a sobreado mais escuro. A proporção entre os dois é a probabilidade do atributo variar em pessoas reais e tem de ser estimada empiricamente.

Essa zona de pares verdadeiros com nomes diferentes revela a principal e mais significativa diferença entre o conceito dos pesos binários, tal como foi originalmente formulado, e o resultado da simplificações acima apresentadas que reencontraram o conceito de quantidade de informação. De facto, o numerador da fórmula respectiva, que significa "probabilidade de o atributo coincidir em pares verdadeiros" só se reduz à frequência relativa quando não

há variabilidade. Quando esta existe, como nos exemplos que demos na secção 3.4.5.3, então a proporção deve ser menor e devemos por isso diminuir o valor do numerador.

Temos por isso de estimar a probabilidade de um valor variar dentro de pares verdadeiros. Expressaremos essa probabilidade, como normalmente, por uma percentagem. Por exemplo, podemos estipular que o nome A varia em 10% dos casos de pares verdadeiros. Isso significa que só nove e não dez dos pares verdadeiros em que o nome A ocorre contêm ocorrências simultâneas desse nome. O outro, correspondendo a 10%, terá o nome A numa ocorrência e um nome diferente na outra.

A fórmula final do peso binário ficaria então:

$$PB_A = \log_2\left(\frac{frel_A \bullet (100\% - pVar)}{frel_A^2}\right)$$

(aplicável quando os atributos concordam)

em que pVar é a probabilidade do atributo variar em pares verdadeiros e $frel_A$ a frequência relativa do atributo A·

Esta é a fórmula que utilizamos. A questão que resta é a seguinte: como estimamos pVar na fórmula acima? Aqui, na verdade, não há simplificação possível. É necessário analisar um conjunto de pessoas já identificadas para obter estas probabilidades. Os cálculos podem ser refeitos periodicamente à medida que a identificação vai progredindo. A actual implementação do sistema assume pVar perto de zero quando não há informação disponível. Em qualquer momento o utilizador pode desencadear um recálculo das probabilidades de variação a partir das pessoas nesse momento identificadas.

Nem sempre há informação disponível para calcular a probabilidade de um atributo variar. Para isso basta que, num dado momento, não exista um conjunto significativo de pessoas identificadas em que esse atributo esteja presente. O modo de proceder nesse caso é o seguinte: se não existe informação disponível sobre a probabilidade de determinado atributo variar, utiliza-se a probabilidade geral para atributos desse tipo. Por exemplo, se não tivermos informação sobre a probabilidade de determinado local de residência variar então utilizamos a probabilidade de a residência em geral variar. O sistema mantém por isso duas séries de estimativas em simultâneo: uma estima a probabilidade de variar cada valor específico de cada atributo (cada profissão, cada local, cada nome); outra estima a probabilidade do atributo variar independentemente do valor concreto que assume na pessoa em curso de identificação.

O modelo dos pesos binários é igualmente usado para quantificar a discordância. Quando comparamos dois atributos de duas ocorrências nominiais e estes concordam usamos a fórmula anteriormente explicada para quantificar a semelhança. Quando os atributos não concordam, tendo por isso valores diferentes, quantificamos igualmente o efeito dessa discordância na construção de uma conclusão sobre se de facto as pessoas são as mesmas ou não. Ao contabilizarmos as discordâncias criamos um processo **aditivo** no qual os vários atributos são comparados, com aqueles que concordam a contribuir positivamente para a solução final e os que discordam a contribuir negativamente.

Como quantificamos as discordâncias? Exactamente do mesmo modo. Se, quando um atributo é idêntico, calculamos a razão entre a frequência com que o atributo concorda em pares verdadeiros e a frequência correspondente nos pares falsos então, quando os atributos discordam, calculamos a razão entre a

frequência com que o atributo discorda em pares verdadeiros e a frequência correspondente nos pares não verdadeiros.

No fundo o raciocínio é este: o facto do atributo discordar deve contribuir negativamente para a nossa certeza de que o par diz respeito à mesma pessoa. Como é evidente, os atributos discordam mais frequentemente entre ocorrências de pessoas diferentes do que entre ocorrências da mesma pessoa. Assim a razão entre as discordâncias em pares verdadeiros e as discordâncias em pares falsos tende sempre a ser menor que a unidade, fornecendo um valor negativo para o cálculo do peso binário.

Utilizando as simplificações acima podemos construir a fórmula a aplicar quando um atributo discorda:

$$PB_A = \log_2\left(\frac{frel_A \bullet pVar}{frel_A - frel_A^2}\right)$$

(aplicável quando os atributos discordam)

O numerador desta fração indica a probabilidade do atributo discordar em pares verdadeiros. Assim, por exemplo, se a probabilidade do atributo variar em pares verdadeiros é 10% e se a probabilidade do atributo ocorrer (frel) é 12.5% o numerador fica igual a 1,25%. No fundo estamos a contabilizar a importância da zona de sombreado claro no diagrama 3.9.

O denominador é dado pela probabilidade do atributo discordar entre os pares falsos. Como vimos, podemos assumir o total de pares como uma aproximação aceitável do total de pares falsos (ver acima diagrama 3.8). Do total de pares falsos o atributo ocorre numa proporção que é dada pela frequência relativa. Dessa proporção uma parte corresponde a pares em que o atributo concorda. Essa parte tende para frequência relativa ao quadrado (a zona a tracejado no diagrama 3.10). Assim a probabiblidade de um atributo

discordar em pares falsos pode ser aproximada pela diferença entre a frequência relativa e a frequência relativa ao quadrado, que aqui representa a diferença entre o total de pares em que o atributo ocorre menos aqueles pares em que o atributo concorda (ver diagrama 3.10).

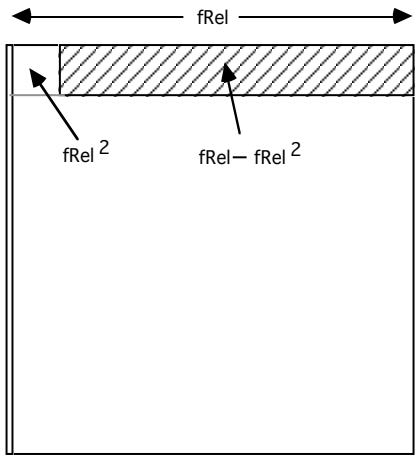


Diagrama 3.10: a frequência com que um atributo discorda em pares falsos é calculada a partir da frequência relativa em que o atributo ocorre dimuída da parte correspondente aos pares em que o atributo concorda.

E com esta justificação encerramos a derivação das fórmulas que se utilizam para quantificar a semelhança entre ocorrências nominais.

Este modelo permitiu um conjunto de simplificações que grandemente facilita a definição dos aspectos matemáticos do cruzamento nominal. Como se vê as fórmulas finais de valorização das semelhanças e diferenças assentam apenas em dois valores: o da frequência relativa, que se assume seja representativa da população real e a probabilidade do atributo variar, que é calculada por amostragem a partir de pessoas já identifficadas. Encerraremos esta secção com algumas observações sobre a importância relativa destes valores e o modo como interagem em situações reais.

Quadro 3.34: Cálculo de pesos binários em situações reais: nas tabelas seguintes vê-se que os pesos binários dependem fortemente da frequência relativa quando os atributos concordam. Quando o atributos discordam a probabilidade de variarem em pessoas reais é usada para relativar o impacto negativo no peso binário. Por outro lado mesmo taxas de erro elevadas na estimação de pvar revelam pouco impacto nos resultados finais

			Concordantes			Discordantes		
Atributo	*pvar*	*frel*	*P(A)*	*P(B)*	*PB*	*P(A)*	*P(B)*	*PB*
Apelido								
rodrigues	5.1%	16.5%	15.7%	2.7%	**2.5**	0.84%	13.77%	**-4.0**
vasconcelos	44.2%	0.2%	0.1%	0.0003%	**8.4**	0.07%	0.16%	**-1.2**
Residencia								
soure	32.2%	16.0%	10.9%	2.6%	**2.1**	5.16%	13.46%	**-1.4**
qta s.tome	29.8%	0.1%	0.1%	0.00009%	**9.5**	0.03%	0.10%	**-1.7**

Mesmos dados com variações artificias de 15% em pvar

			Concordantes			Discordantes		
Atributo	*pvar*	*frel*	*P(A)*	*P(B)*	*PB*	*P(A)*	*P(B)*	*PB*
Apelido								
rodrigues	4.3%	16.5%	15.8%	2.7%	**2.5**	0.71%	13.77%	**-4.3**
vasconcelos	50.8%	0.2%	0.1%	0.0003%	**8.2**	0.08%	0.16%	**-1.0**
Residencia								
soure	27.3%	16.0%	11.7%	2.6%	**2.2**	4.39%	13.46%	**-1.6**
qta s.tome	34.2%	0.1%	0.1%	0.00009%	**9.4**	0.03%	0.10%	**-1.5**

O quadro 3.34 mostra o resultado da aplicação das fórmulas derivadas anteriormente a dados reais. O objectivo destes números é demonstrar duas coisas: que o papel de pVar (a probabilidade de um atributo variar em pessoas reais) é relativamente pouco importante face à frequência relativa de cada atributo. Esse papel menor revela-se em dois aspectos principais: em primeiro lugar, valores diferentes de pVar afectam pouco o peso binário positivo de um

atributo; em segundo lugar, e como consequência disso, variações em pVar introduzidas por imprefeições no seu processo de cálculo não afectam significativamente os resultados.

O primeiro aspecto a salientar é o da relativa neutralidade de pVar em relação aos pesos binários positivos. Veja-se o peso binário associado ao apelido "Rodrigues" e o peso binário associado à residência "Soure". Ambos os atributos têm uma frequência relativa aproximada, à volta de 16%. Contudo o apelido "Rodrigues" tem um probabilidade de variar seis vezes menor que a residência "Soure"[52]. No cálculo do peso binário positivo, em caso de concordância os valores, os valores encontrados são próximos (2.5 para o apelido e 2.1 para a residência).

Compare-se o que acontece quando a residência é a quinta de S.Tomé. A frequência relativa é bastante inferior a Soure embora a probabilidade de variar seja bastante semelhante (29.8% para 32.2%). A diferença nos pesos binários é assinalável (9.5 para 2.1). A frequência relativa determina assim muito mais fortemente o resultado que pVar.

Qual então o papel de pVar? Esse papel revela-se sobretudo no cálculo do peso binário em caso de discordância. O efeito detecta-se melhor na comparação entre um nome comum mas estável, como "Rodrigues" e um nome raro mas instável, como "Vasconcelos". O apelido "Vasconcelos" tem um probabilidade de variar muito alta, 44%. Esse valor decorre do facto do apelido "Vasconcelos" aparecer como parte de nomes longos, como "João da Costa

[52] O valor de 32.2% para a probabilidade de a residência "Soure" variar em pessoas reais está aqui um pouco inflacionado em relação ao que se obteria a partir de uma amostra mais significativa de pessoas identificadas. Contudo é natural que a sede de frequesia tenha associada uma taxa de variabilidade importante porque em muitas fontes a residência pode ser transcrita simplesmente como a paróquia. É o caso de documentação de origem central. Em fontes locais as mesmas pessoas são referidas com residência mais precisa. Daí a variabilidade grande associada a este atributo.

Cabral e Vasconcelos", que são frequentemente abreviados, por exemplo para "João da Costa". Assim é comum que o apelido Vasconcelos não esteja simultaneamente presente em duas ocorrências da mesma pessoa real. Por outro lado, como é um atributo raro, o número de pessoas reais que podemos utilizar para estimar a probabilidade de variação é muito reduzido o que faz com que algumas variações tenham um efeito grande no valor final. Por todas estas razões o valor de pVar para o apelido "Vasconcelos" é grande. A fórmula capta esse facto de maneira eficaz: o peso binário negativo é de apenas -1.2. Isso significa, na prática, que a penalização para o atributo Vasconcelos estar presente em apenas uma de duas pessoas que estamos a comparar é de -1.2. Mas a vantagem de estar presente em ambas é de 8.4 pontos.

Ao mesmo tempo podemos ver que a sensibilidade geral dos resultados a variações em pVar é pequena. O segundo quadro da tabela 3.34 mosta o resultado de introduzir variações de 15% nos valores de pvar. Como se vê o efeito sobre o resultado final é muito pequeno, mesmo tendo havido o cuidado de introduzir variações divergentes. O significado deste teste é importante porque o processo pelo qual chegamos aos valores de pvar é necessariamente frágil e sujeito às contingências da amostragem. Assim, podemos concluir que mesmo que o método de amostragem que utilizamos para chegar a pVar seja imperfeito a influência dessa limitação nos resultados finais é pequena.

Em conclusão, vemos que a fórmula de cálculo dos pesos binários parece fornecer uma modelização bastante eficaz do processo de valorização da informação. Alguns dos valores presentes nessa fórmula são mais determinantes que outros. O modelo tem uma "robustez" assinalável, ou, por outras palavras, é bastante imune a variações pequenas nos parâmetros que o

regem[53]. O valor crítico aqui é a frequência relativa, complementada por uma aproximação, ainda que grosseira, da probabilidade de variação de cada atributo.

Ao concluir esta secção pensamos ser útil fazer uma recapitulação dos principais conceitos introduzidos. Do ponto de vista formal aqui ficou explicitado o modelo probabilístico que serve de base à quantificação da semelhança de ocorrências de pessoas em fontes. O modelo utilizado não se limita a apresentar os princípios estatísticos relevantes para este tipo de situações. Procurou-se fornecer uma *tradução operacional* dos princípios formais que, através de simplificações assumidas, permitissem uma implementação clara e eficaz. Para isso servimo-nos de uma simulação que, reduzindo o sistema a um esquematismo compreensível, permitisse visualizar e entender o modo como interagem as várias componentes do modelo.

O passo seguinte é entender como combinamos os pesos calculados para os vários atributos e ver alguns exemplos dos resultados obtidos. Isso permitir-nos-á abordar a questão dos limiares de aceitabilidade, um tópico central na literatura sobre esta matéria.

---

[53] Esta constatação vai ao encontro das conclusões obtidas por outros investigadores, e é um dos argumentos fortes a favor da utilização de probabilidades condicionais na modelização de conhecimento: " *[...] Bayesian models tolerate large devitions in the prior and conditional probabilites. That is, even rough estimates for which qualitative expressions such as 'rare', 'frequent', and 'probable' serve as guidelines may be accurate enough to result in the recommendation of the correct decision*", Ben-Basset, M., Klove, K.L., Weil, M.H., *Sensisivity Analysis in Bayesian Classification Models: Multiplicative Deviations* . In: "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. PAMI-2, Nº3, citado por Neapolitan, Richard, *ob. cit.*, p.74.

3.4.4.5. Aditividade e limiares.

Na secção anterior vimos como se constrói o cálculo do peso binário na comparação de dois atributos. Quando comparamos duas ocorrências de pessoas em fontes temos vários atributos disponíveis. Nesta secção iremos ver como se combinam os pesos dos vários atributos e se utiliza o valor do resultado final.

Uma das principais características do peso binário binário é a sua natureza aditiva. Essa é, aliás, a razão pela qual se utiliza o logaritmo de base 2 no seu cálculo: para ter um valor adicionável.

Com vimos, em caso de discordância, a fórmula do peso binário produz um valor negativo. Assim, comparando sucessivamento os vários atributos em presença e adicionando os pesos binários calculados obtemos um valor global que representa o *valor da ligação* entre duas ocorrências, conceito que introduzimos antecipadamente (ver supra, secção3.4.2).

Um aspecto amplamente discutido tem sido o de fixar o limiar a partir do qual o peso da ligação é aceitável e abaixo do qual abandonamos o resultado da comparação. É impossível termos uma ideia a priori de qual deva ser esse limiar. Qualquer decisão nessa matéria tem de ser suficientemente flexível para prever, à partida, que haverá ligações correctas abaixo do limiar, e que, inversamente haverá situações em que ligações acima do limiar terão que ser abandonadas por incorrectas.

No desenho deste sistema pretendeu-se manter as escolhas mínimas, tornando-se o modelo tão autónomo quanto possível. Contudo a escolha do limiar não está automatizada. Ela deverá ser escolhida pelo utilizador e é uma escolha importante porque influenciará o comportamento global do sistema. Um limiar baixo provocará um comportamente mais "optimista" e tendencialmente o sistema estará mais sujeito a aceitar erroneamente duas

pessoas diferentes como sendo a mesma. Um limiar mais alto provocará atitudes mais "pessimistas" e é possivel que os programas de cruzamento separarem ocorrências diferentes da mesma pessoa real. A nossa escolha de limiar foi de 11.5 que, no nossos dados, provoca um comportamento ligeiramente "pessimista".

A ideia de limiar surge naturalmente no seguimento da utilização do peso binário. Se podemos quantificar a comparação entre os atributos de duas pessoas então deverá haver um ponto que separa as pessoas que de facto são as mesma das que o não são. A experiência demonstra que de facto não é assim. Quando se examina os resultados das comparações e da valorização feita pela adição de pesos binários chega-se à conclusão que há uma zona de sobreposição entre os valores das comparações entre pessoas reais e pessoas falsamente identificadas (diagrama 3.11).

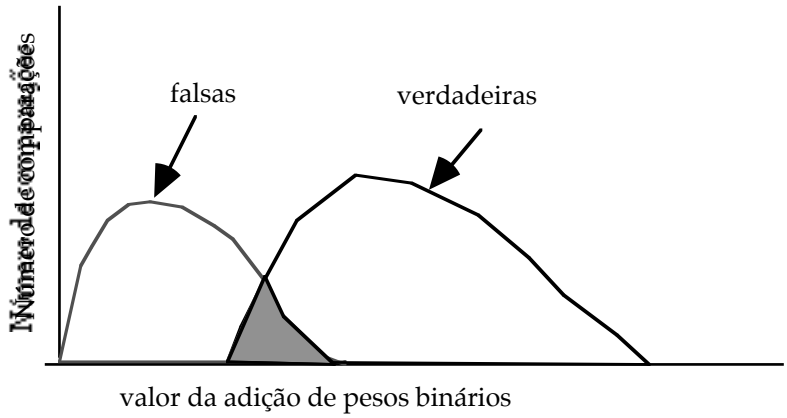


Diagrama 3.11: Ao classificarmos comparações entre ocorrências de pessoas conforme dizem respeito a pessoas reais ou a falsas identificações é notório que não há um limiar pontual onde se possam separar os dois tipos. Antes há uma zona de sobreposição onde encontramos tanto ligações falsas como verdadeiras

Alguns autores utilizam um duplo limiar: se o valor da comparação dos atributos de duas ocorrênias de pessoas é superior a determinado valor, chamado de aceitação, então a identificação é retida, se fôr inferior a um segundo valor, dito de rejeição, a identificação não é feita, isto é, conclui-se claramente que duas pessoas não são a mesma. Na terceira hipótese, em que o valor se encontra entre o limiar de rejeição e o limiar de aceitação temos uma situação em que não é possível tomar uma decisão. Procura-se então o melhor processo de encurtar a diferença entre os dois valores, minimizando a intervenção humana[54].

O conceito de duplo limar pressupõe que de algum modo se pode determinar se as ligações são verdadeiras ou não de forma independente do próprio cálculo dos pesos binários. Normalmente o resultado de uma identificação feita com base nessa medida é revista manualmente. Assim se detectam as ligações falsas e verdadeiras e se pode construir um gráfico como o apresentado acima (diagrama 3.11). Isso significa que existe outro tipo de conhecimento relevante para o processo de identificação que não é colhido pelo formalismo probabilístico apresentado. Parte desse conhecimento pode basear-se em informação que não está disponível ao sistema, ou que seria demasiadamente custoso, em termos computacionais, considerar. Por exemplo, podemos ter assinaturas comparáveis que nos fornecem pistas de identificação

---

[54] Ver Fellegi,Ivan; Sunter Alan: *A Theory for Record Linking* . Journal of the American Statistical Association, nº64, p.1183-1210, 1969. A utilização intensiva de notação matemática neste artigo torna a sua leitura algo inacessível. Ver um resumo em Winchester, Ian - OA brief survey of the algoritmic, mathematical and philosophical literature relevant to historical record linkage" in Wrigley, E.A. (ed), *Identifying People in the Past.* pp. 142 e ss. Os sistemas controlados por duplo limiar são normalmente sistemas que trabalham com dois ficheiros não redundantes e procuram optimizar a escolha do melhor par. No nosso caso, como veremos, existe um limiar de referência e um duplo limiar dinâmicamente estimado para cada grafo de ligações.

que os programas não têm acesso. mas muitas vezes as comparações feitas pelo método dos pesos binários são validadas em termos de coerência global das soluções que implicam. Por outras palavras: existem constrangimentos que guiam a identificação que só são visíveis quando juntamos o resultado de todas as aproximações feitas pelo método dos pesos binários e avaliamos a biografia ou conjunto de biografias como um todo. Parte desse processo é automatizável também e tem a ver, como referimos no início, com a construção dos grafos de ligações e o seu tratamento.

A solução que aqui propomos automatiza parte dessa validação a posteriori das identificações feitas. O modo como isso é feito tem a ver com o processo de resolução de ambiguidades. Em termos gerais podemos dizer que os algoritmos utilizados têm a capacidade de considerar falsa uma ligação acima do limiar pré-definido e considerar verdadeira uma ligação abaixo desse limiar e que o fazem automaticamente com base em critérios de coerência. Por outras palavras, se, para manter a coerência geral de uma solução óptima for necessário "forçar" o limiar, os programas têm a capacidade de o fazer automaticamente. É nossa experiência que esta flexibilidade é determinante na eficácia de uma solução automatizada. O limiar, como número "mágico", não existe. Existe sim um indicador genérico do tipo de comportamento que se pretende.

Os procedimentos que inserem o método dos pesos binários num sistema mais alargado que utiliza o conceito de coerência não são excepcionalmente complexos. Antes se apoiam numa representação formalizada de operações de senso comum.

Um aspecto importante desta parte do sistema reside no facto de ser aqui, mais do que em qualquer outro lado, que se joga o sucesso ou o insucesso do

empreendimento. Trata-se também da parte de todo o sistema que mais directamente modeliza um comportamento aparentemente "inteligente", embora o processo de aferição do valor das ligação, que atrás ficou explicado, seja já uma componenete importante desse comportamente.

Não é possível dar uma descrição da solução implementada sem ter uma ideia mais clara do problema. Iremos assim introduzir um exemplo concreto em que partimos de uma referência nominal avulsa e recolhemos todas as ocorrências que podem eventualmente ser a mesma pessoa. Aí poderemos ter uma visão clara do problema que resta após a aplicação da fórmula dos pesos binários a um conjunto de candidatos. Ficamos, como se verá, com uma colecção de referências sobre a qual teremos ainda que decidir se dizem respeito todas à mesma pessoa ou não.

3.4.5. UM EXEMPLO.

Em 13 de Outubro de 1695 casaram-se Mateus Gomes e Maria Rodrigues. É o quarto casamento da nossa série e serve bem de partida para uma exemplificação do modo de procedimento dos algoritmos desenvolvidos para este trabalho. Com este exemplo os problemas finais ficam definidos e poderemos de seguida terminar o capítulo revendo de forma mais formal os aspectos heurísticos da solução produzida.

Em treze de Outubro de 1695 casam-se Mateus Gomes e Maria Rodrigues na igreja matriz de Soure. A transcrição do acto, na linguagem kleio, é a seguinte:

```
casamento$4/1695-10-13/igreja matriz/celebrante=luis alvares pinto
  noivo$mateus gomes/m/id=c4-1
```

```
    pnoivo$manuel gomes/m/id=c4-2
       atr$residencia/paleao
       atr$profissao/ferreiro
    mnoivo$maria simoes/f/id=c4-3
       /obs=comentário ao campo-nome:nao diz "sua mulher" como
habitualmente.
  noiva$maria rodrigues/f/id=c4-4
    pnoiva$inacio vaz/m/id=c4-5
       atr$morto/antes
    mnoiva$maria rodrigues/f/id=c4-6
       atr$residencia/carvalheira de cima
       atr$ec/v
  test$manuel rodrigues/m/id=c4-7
     atr$residencia/carvalheira de cima
     rel$parentesco/irmao/maria rodrigues/c4-4
     rel$parentesco/irmao/diogo rodrigues/c4-8
  test$diogo rodrigues/m/id=c4-8
     atr$residencia/carvalheira de cima
     rel$parentesco/irmao/maria rodrigues/c4-4
     rel$parentesco/irmao/manuel rodrigues/c4-7
```

Supondo que estamos interessados em saber mais sobre o noivo, e pedindo ao sistema outras ocorrências que possam dizer respeito à mesma pessoa, surge o seguinte assento de casamento:

```
casamento$c1717.20/1717-10-28/igreja matriz/celebrante=manuel
rodrigues (vigario)
  noivo$mateus gomes/m/id=c1717.20-p1
    pnoivo$manuel gomes/m/id=c1717.20-p2
       a$residencia/paleao
    mnoivo$maria simoes/f/id=c1717.20-p3
    mulher1$maria rodrigues/f/id=c1717.20-p4[55]
      pmulher1$inacio vaz/m/id=c1717.20-p5
         a$residencia/carvalheira de cima
      mmulher1$isabel rodrigues/f/id=c1717.20-p6
  noiva$maria esteves/f/id=c1717.20-p7
    pnoiva$manuel mateus/m/id=c1717.20-p8
       a$morto/antes
       a$residencia/paleao
    mnoiva$ana esteves/f/id=c1717.20-p9
       a$morta/antes
  test$antonio gomes/m/id=c1717.20-p10
     a$residencia/sobrado
  test$francisco rodrigues gante/m/id=c1717.20-p11
     a$residencia/paleao
  test$manuel rodrigues branco/m/id=c1717.20-p12
     a$residencia/paleao
  test$pedro francisco/m/id=c1717.20-p13
     a$residencia/casconho
```

[55] O grupo "mulher1" num casamento significa "primeira mulher" e usa-se por isso no casamento de viúvos, que é o caso.

Como vamos contabilizar se se trata da mesma pessoa ou não? Em primeiro lugar é necessário referir que, ao comparmos as duas ocorrências de Mateus Gomes, não iremos considerar estritamente os atributos pessoais, que aqui são pouco mais do que o nome. Teremos obviamente que entrar em linha de conta com a informação respeitante aos parentes. Neste sistema as pessoas herdam atributos dos parentes próximos. Assim um dos atributos de ambos os Mateus Gomes será o apelido do pai; outro o nome da mulher. Assim a comparação entre as duas ocorrências processa-se do seguinte modo[56]:

```
pessoaA: c4-1
pessoaB: c1717.20-p1
apelido
   valor A=gomes valorB=gomes
   peso binário = 4.75
   acumulado = 4.75
apelido-mae
   valor A = simoes valorB = simoes
   peso binário = 3.45
   acumulado  = 8.20
apelido-mulher
         valor A = rodrigues valorB = esteves
   peso binário = -3.59
         valor A = rodrigues valorB = rodrigues
   peso binário = 2.47
   acumulado = 10.67
apelido-pai
         valor A = gomes valorB = gomes
   peso binário = 4.58
   acumulado = 15.25
ec (estado civil)
         valor A=c valorB=c
   peso binário = 0.32
      valor A=c valorB=v
   peso binário = -4.05
   acumulado = 15.58
pnome
         valor A=mateus valorB=mateus
   peso binário = 6.125
   acumulado = 21.708
pnome-mae
         valor A=maria valorB=maria
   peso binário= 1.27
   acumulado = 22.97
pnome-mulher
```

[56] O que se segue é uma listagem fornecida pelo sistema, depois de simplificada e formatada.

```
          valor A=maria valorB=maria
   peso binário = 1.31
   acumulado = 24.28
pnome-pai
          valor A=manuel valorB=manuel
   peso binário = 1.5
final = 25.8
```

Salientemos alguns aspectos menos evidentes. A comparação dos atributos, aqui apresentada por ordem alfabética, é neste caso sempre acumulativa, isto é, para todos os atributos presentes foi possível encontrar sempre uma concordância. Em dois atributos, "apelido da mulher" e "estado civil" temos valores que não concordam ao mesmo tempo que temos valores que concordam. O casamento de 1717 fornece dois nomes de mulheres para Mateus Gomes, uma vez que se trata de um recasamento. Acontece que o primeiro nome de ambas é o mesmo: "Maria". O apelido é diferente "Rodrigues" e "Esteves". Ao comparar as duas ocorrências o sistema aponta a falha de concordância do apelido da mulher, que penalizaria o total em -3.59 pontos. Contudo, como consegue obter uma concordância com "Rodrigues", presente em ambos os actos, considera o atributo como concordante e ignora a discordância. Uma decisão semelhante é tomada em relação ao estado civil. Sempre que o sistema encontra vários valores para um mesmo atributo considera que há concordância se pelo menos um par de valores concordar. Se vários concordarem acumula os pesos binários respectivos ignorando os que eventualmente forem diferentes. Só se todos os valores discordarem é que o peso binário negativo é retirado do total acumulado.

Este mecanismo é muito eficaz no tratamento dos nomes, a respeito dos quais por vezes, se inventam fórmulas complexas para lidar com o facto do mesmo nome aparecer com várias partículas e combinação diferentes, como é o caso aqui. Em vez de um processamento complicado simplesmente consideramos que o nome é composto de vários atributos. O atributo "nome" e zero ou várias ocorrências do atributo "apelido". Quando as pessoas têm vários

apelidos, e todos estão presentes em duas ocorrências diferentes da mesma pessoa, então somamos os pesos binários das várias concordâncias. Pensamos que este processo espelha bem o raciocínio natural de identificação de nomes.

Note-se ainda que a comparação se limita aos atributos que Mateus Gomes "herdou" dos familiares mais próximos (pai,mãe,mulher). Poderíamos ter levado a herança mais longe e considerar o "nome do sogro", "apelido da sogra", etc... Os casos em que isso fazia uma diferença são raros e a propagação de heranças em segundo e terceiro grau sobrecarregaria demasiado o sistema pelo que nos circunscrevemos aos parentes directos na propagação dos atributos[57].

Temos, em conclusão uma concordância de 25 pontos. Como vemos pelos dois assentos não restam quaisquer dúvidas que se trata da mesma pessoa.

O sistema fornece uma segunda ocorrência que compara inequivocamente com o noivo do casamento de 13 de Outubro de 1695, e que corresponde ao óbito da primeira mulher, Maria Rodrigues:

```
obito$obitos 1696-o73/1696-12-22/igreja de s.joao
   /padre=luis alvares pinto
   /oficios=3 de 3
   /sacram=sim
  n$maria rodrigues/f/id=obitos 1696-o73-p1
    mr$mateus gomes/m/id=obitos 1696-o73-p2
       a$residencia/paleao
```

O peso da ligação com o Mateus Gomes original é de 14.7 pontos.

---

[57] O principal problema em tornar complexo o mecanismo da propagação de atributos entre pessoas relacionadas é que dificulta as correcções em caso de erros de registo detectados depois da propagação dos atributos estar feita. Aí é necessário propagar a correcção pelas pessoas que "herdaram" o valor erróneo, o que representa sempre uma sobrecarga qualquer que seja o mecanismo concreto pelo qual essa tarefa é executada. Aqui, como em muitos outros pontos do sistema, foi necessário encontrar um equilíbrio.

Ao pesquisar a base de dados para encontrar estas duas ocorrências o sistema comparou os atributos dos vários Mateus Gomes que se encontram registados. Efectuou assim mais de três dezenas de comparações com outras ocorrências. Para cada uma delas somou os pesos binários dos atributos concordantes e subtraiu quando havia discordância. As duas ocorrências que já referimos foram as duas mais pontuadas, mas existem comparações com resultados mais baixos. Vejamos por exemplo este baptismo:

```
baptismo$b1719.91/1719-10-1/celebrante=manuel rodrigues (vigario)
  n$jose/m/id=b1719.91-p1
    pai$manuel gomes ferreiro/m/id=b1719.91-p2
       a$residencia/paleao
       a$naturalidade/paleao
    mae$isabel tome/f/id=b1719.91-p3
       a$naturalidade/casal da tojeira
    pad$jose gaspar/m/id=bpad01101719
       a$ec/s
       a$residencia/paleao
    mad$maria esteves/f/id=b1719.91-p4
      mrmad$mateus gomes/m/id=b1719.91-p5
         a$residencia/paleao
  referido$joao gaspar/m/id=br01101719
     /obs=campo adicional-eid:br01101719.
     a$residencia/paleao
     r$sociabilidade/procurador em baptismo/jose gaspar/bpad01101719
```

Ao comparar esta ocorrência de Mateus Gomes com a do casamento incial temos nome e residência iguais e nome da mulher diferente. O total dos pesos binários é 8.9 pontos. Sabemos, como é evidente, que se trata da mesma pessoa, que casou em 1695 com Maria Rodrigues, enviuvou em 1696 e recasou com Maria Esteves em 1717. O problema é que se considerarmos 8.9 pontos como um valor aceitável para a comparação de duas ocorrências estamos implicitamente a admitir que todos os Mateus Gomes do Paleão são a mesma pessoa, ainda que as mulheres respectivas tenham nomes diferentes.

Para complicarmos um pouco mais vejamos ainda este casamento, em que outro Mateus Gomes, viúvo, se casa com Isabel Rodrigues.

```
casamento$c1720.24/1720-11-24/igreja matriz/celebrante=manuel
rodrigues (vigario)
  noivo$mateus gomes/m/id=c1720.24-p1
    pnoivo$mateus gomes/m/id=c1720.24-p2
       a$residencia/sobral
       a$morto/antes
    mnoivo$margarida lopes/f/id=c1720.24-p3
       a$morta/antes
    mulher1$maria luis/f/id=c1720.24-p4
      pmulher1$bartolomeu luis/m/id=c1720.24-p5
         a$residencia/soure
      mmulher1$gracia gomes/f/id=c1720.24-p6
    mulher2$maria ferreira/f/id=c1720.24-p7
      pmulher2$salvador ferreira/m/id=c1720.24-p8
         a$residencia/soure
         a$morto/antes
      mmulher2$maria simoes/f/id=c1720.24-p9
    mulher3$antonia domingues/f/id=c1720.24-p10
      pmulher3$antonio domingues/m/id=c1720.24-p11
         a$residencia/mamarosa, freguesia de cantanhede
         a$freguesia/cantanhede
         a$morto/antes
      mmulher3$maria domingues/f/id=c1720.24-p12
         a$morta/antes
         a$residencia/mamarosa, freguesia de cantanhede
         a$freguesia/cantanhede
  noiva$isabel rodrigues/f/id=c1720.24-p13
    pnoiva$antonio da gante/m/id=c1720.24-p14
       a$residencia/casais de s.mateus
    mnoiva$maria rodrigues/f/id=c1720.24-p15
       a$morta/antes
  test$joao henriques/m/id=c1720.24-p16
     a$residencia/soure
  test$antonio alvares/m/id=c1720.24-p17
     a$residencia/soure
  test$paulo da gante/m/id=c1720.24-p18
     a$residencia/casais de s.mateus
  test$antonio rodrigues feijao/m/id=c1720.24-p19
     a$residencia/casais de s.mateus
```

Tudo indica que este seja outra pessoa. Apesar de ser um viúvo a recasar pela quarta vez, nenhuma das mulheres de este Mateus Gomes parece adequar-se à informação que possuímos sobre o "nosso" Mateus Gomes. Contudo, e isso é que é nos interessa agora, do ponto de vista da comparação de atributos, existem muitas semelhanças: o mesmo nome, que não é muito comum, o estado civil de viúvo e mulheres com apelido Rodrigues e primeiro nome Maria. O

total, embora inflacionado por todos estes recasamentos, é de 14.6 pontos entre o noivo do casamento acima, de 1720 e o marido da defunta no óbito de 1696[58].

Em conclusão, pela simples comparação dos atributos e pelo cálculo do peso das ligações poderemos obter ligações fracas, que dizem respeito à mesma pessoa real com é o caso de:

Manuel Gomes, casado com Maria Rodrigues (casamento de 1695) e Manuel Gomes do Paleão casado com Maria Esteves (8 pontos).

Ao mesmo tempo podemos obter ligações mais fortes que aproximam pessoas que parecem ser diferentes:

Manuel Gomes viúvo de Maria Rodrigues e Manuel Gomes viúvo de Maria Domingues (14 pontos).

Esta dificuldade em acertar um limiar acima do qual as ligações sejam verdadeiras e abaixo do qual sejam falsas é uma característica muito importante do formalismo do cálculo binário. É igualmente evidente que só sabemos que, no primeiro caso acima, se trata da mesma pessoa porque temos o casamento de 1717 que faz a ponte entre as duas esposas de Manuel Gomes. Por outras palavras, a validade da ligação não depende exclusivamente das duas ocorrências mas sim de uma terceira que valida indirectamente outra.

---

[58] O sistema considera que o primeiro nome e apelido da mulher (Maria Rodrigues) concordam nos dois actos, apesar de, no casamento de 1720, o apelido "Rodrigues" se encontrar numa das esposas (Isabel Rodrigues) e "Maria" noutras duas confusão compreensível dado o número de mulhers que Mateus Gomes, noivo em 1720, teve. Por razões demasiadamente técnicas para expôr aqui é complicado evitar que o sistema use primeiros nomes e apelidos de diferentes parentes do mesmo tipo, caso que aliás só se aplica aos recasamentos, e, como veremos, é possível corrigir automaticamente estas confusões a posteriori não foi considerado justificável tornar mais complexos os procedimentos. De qualquer modo, mesmo que descontemos os 1.3 pontos do primeiro nome "Maria" o valor de concordância continua a ser alto o que se justifica pelo facto de não considerarmos determinável a residência de um noivo num casamento, porque, por definição, se encontra potencialmente entre duas casas.

Estes aspectos convergem na necessidade de complementar o cálculo dos pesos binários por mecanismos que permitem a flexibilidade necessária para não raciocinar exclusivamente em termos de limiar de aceitabilidade, nem tomar decisões baseadas apenas na comparação de duas ocorrências.

O primeira conclusão diz respeito ao modo como são gerados os candidatos. Como vimos acima, se partirmos do casamento de 1695, recolhemos dois candidatos bem valorizados: o óbito da primeira mulher e o recasamento em 1717. Só após recolhermos o recasamento é que podemos avaliar o interesse do baptismo de 1719 em que Mateus Gomes acompanha a segunda mulher enquanto madrinha. Significa isto que cada nova ocorrência que aparece obriga a relançar a pesquisa e reavaliar candidatos anteriormente rejeitados. Este processo só se esgota quando todos os candidatos aceites tiverem provocado a respectiva reavaliação. Tal como um gota que cai na superfície de um líquido, o processo de identificação vai se progagando pelo espaço nominal até esgotar todas as alternativas.

Assim, ao lançarmos o sistema à procura de todas as ocorrências do Mateus Gomes que se casa com Maria Rodrigues em 1695 serão recolhidas ao todo 22 ocorrências, que, como veremos, dizem respeito a 4 pessoas reais diferentes. Partindo da ocorrência incial recolhe-se o óbito da primeira mulher e o recasamento do viúvo. Partindo do recasamento do viúvo recolhe-se mais o baptismo em que é madrinha a segunda mulher, mas começam a agregar-se também ocorrências menos claras como o recasamento de um Mateus Gomes em 1720 e o óbito de um viúvo, Mateus Gomes, em 1687 (este será abandonado por inconsistência cronológica como veremos adiantes). A partir do casamento de 1720 serão recolhidas mais dez ocorrências de Mateus Gomes na base de

dados. Quando o processo de geração terminar temos as 22 ocorrências recolhidas.

Como vimos nos exemplos que demos, e na sequência das considerações sobre limiares anteriormente feitas, sabemos que entre estas 22 ocorrências existem ligações verdadeiras e outras falsas, sem que haja um valor claramente distintivo entre os dois conjuntos. Por outro lado, além das variações da soma dos pesos binários, existem também ligações que são falsas por questões de lógica e não de probabilidades. É o caso das ligações entre um óbito de uma pessoa e ocorrências posteriores no tempo.

Para podermos apreender a interacção de todos estes factores é indispensável visualizar graficamente o espaço das ligações (ver figuras extra-texto nº 23 e 24) . Em apêndice fornecemos todos os actos envolvidos neste exemplo. A visualização do grafo fornece quase por si só a solução para a identificação das pessoas reais envolvidas nestas 22 referências[59].

O gráfico posiciona as ocorrências de maneira a que, grosso modo, a semelhança calculada pelos pesos binários tenha uma correspondência na distância entre as legendas respectivas no plano. Com efeito, a representação gráfica procura optimizar a correlação entre a distância no plano e a "proximidade probabilística" calculada pelo método dos pesos binários. Embora não exista necessariamente uma correspondência exacta entre os dois níveis, e o gráfico seja por isso sempre uma aproximação, ou melhor, uma

[59] O programa que desenha os gráficos a partir dos grafos de cruzamento nominal foi especialmente desenvolvido pelo autor para este projecto, a partir da quase impossibilidade de avaliar o espaço do problema examinando listagens de dezenas de ligações. O algoritmo original foi obtido em Michelet, Bertrand-*L'analyse des associations*., Paris, CDST/CNRS, 1987, apêndice B. Revelou-se uma ferramenta de extrema utilidade, transcendendo a sua aplicação original. Com efeito o programa transforma qualquer rede num gráfico de modo que as características estruturais dos dados se tornam visualmente evidentes.

projecção, a experiência demonstra que a visualização dos grafos do cruzamento nominal é um factor determinante na compreensão do processo.

Cada ocorrência é representada por uma etiqueta com a respectiva matrícula de identificação[60]. As ligações são representadas por linhas unindo as várias etiquetas. As linhas são coloridas incialmente com duas côres, indicando se o valor da ligação se encontra acima ou abaixo do limiar de aceitabilidade definido. Neste exemplo o azul escuro define uma ligação "forte" e o verde claro uma ligação "fraca". O valor do limiar, neste exemplo, foi de 11.5.

Este grafo foi gerado a partir da ocorrência incial, "c4-1" que corresponde, como vimos acima, ao casamento de Mateus Gomes com Maria Rodrigues. O processo pelo qual o grafo é criado está exemplificado na figura 23. A partir da ocorrência inicial são recolhidos dois candidatos "c1717.20-p1", que corresponde ao recasamento de Mateus Gomes com Maria Esteves, e "obitos 1696-o73-p2", o óbito da primeira mulher, Maria Rodrigues. Em toda a base estes foram os dois únicos candidatos que, comparados com a ocorrência original, ultrapassaram o limiar de semelhança.

O sistema recomeça a expansão a partir destes dois novos candidatos. Expandindo a partir de "c1717.20-p1" encontram-se mais duas ocorrências. Uma delas é o baptismo em que a segunda mulher de Mateus Gomes, Maria Esteves, aparece como madrinha acompanhada pelo marido[61]. Note-se que esta referência não era "alcançável" directamente a partir da ocorrência original. Foi

---

[60] Sobre as matrículas de identificação ver acima 3.2.2. Nos actos transcritos no apêndice as matrículas de identificação são mostradas a seguir a cada ocorrência nominal.

[61] "Acompanhada" no sentido em que ambos estão presentes no registo do acto. As mulheres são sempre identificadas em relação a um homem, o marido ou o pai. É raro aparecer o nome da mulher isoladamente, e as excepções são significativas, decorrendo da importância das pessoas em questão dentro da comunidade. Neste caso é possível que Mateus Gomes estivesse presente na cerimónia mas não é certo.

necessária a informação constante no recasamento de 1717 para permitir a identificação.

Um dos problemas que se tem de enfrentar é que, por este processo, acontece agregarem-se pessoas que nada têm a ver com o ponto de partida. Assim, nos passos seguintes são associadas ocorrências progressivamente mais distantes da pessoa inicial, mas inter-ligadas por sucessivas relações fortes. No passo 5 da figura 23 vê-se ser associado o casamento de Mateus Gomes com Isabel Rodrigues em 1720 (ocorrência c1720.24-p1). Como este casamento inclui referências às três mulheres anteriores de Mateus Gomes, contém uma quantidade de informação muito grande. Isso provoca uma avalanche de candidatos adicionais, com os baptismos e óbitos dos numerosos familiares deste segundo Mateus Gomes a serem incluídos no grafo[62].

Por cada nó do grafo o sistema reinterroga a base de dados e pesquisa todas as ocorrências com atributos semelhantes[63]. Isso significa que muitas referências nominais são reavaliadas repetidamente, cada vez por comparação com uma nova pessoa. Este processo é necessariamente moroso e consome a parte mais significativa do tempo de identificação numa base de dados que contém perto de 35.000 referências nominais. Poder-se-ia ser tentado a restringir esta procura mas o modo como funciona o sistema torna preferível a sobre-geração do que o contrário, uma vez que é mais fácil detectar eventuais erros na fase de resolução de ambiguidades do que "repescar" uma referência que foi perdida durante a fase de geração.

---

[62] Vejam-se os actos transcritos em apêndice. Este exemplo, curiosos a vários títulos, mostra adicionalmente como um acto tardio faz luz sobre um grande número de actos anteriores. Mais uma razão para construir sistemas abertos que permitem relançar a identificação em qualquer momento.

[63] A estratégia exacta de pesquisa utilizada na fase de geração será mais detalhadamente explicada adiantes

Cada nova expansão a partir de um nó do grafo pode trazer, ou não, novos candidatos. Na maior parte dos casos a re-expansão torna a gerar ocorrências já recolhidas. Chegamos assim a um ponto em que expandimos o último candidato recolhido, não se encontrando mais referências novas. Aí a expansão pára e a fase de geração termina.

Terminada a fase de geração o sistema tem de tentar determinar quantas pessoas reais estão incluídas neste grafo. Embora na descrição que estamos a fornecer do processo utilizemos o conhecimento posterior de quem é quem, o computador, neste momento, tudo o que tem é uma rede de ligações de valores vários, umas mais fortes, outras mais fracas, e é com base nessa representação do problema que se detectará a solução final.

A primeira acção a seguir à expansão do grafo é a aplicação de regras de consistência lógica entre as referências recolhidas. Essas regras exprimem um conhecimento sobre os constrangimentos que ocorrem nas pessoas reais que é independente das considerações probabilísticas que temos vindo a abordar. Por exemplo, sabemos que não podemos ter dois óbitos numa mesma pessoa, nem dois baptismos. Sabemos também que se temos um óbito entre as ocorrências do grafo então algumas ocorrências posteriores no tempo devem dizer respeito a diferentes pessoas reais.

Neste caso temos um óbito em 1687 que foi associado ao grafo no passo 4 (ver figura 23). Se um Mateus Gomes morreu em 1687 seguramente não é a mesma pessoa que casou em 1694, 1695, 1703, 1717 e 1720. De modo que qualquer solução que se encontre não poderá incluir na mesma pessoa real estas referências. As ligações entre as ocorrências respectivas têm por isso de ser marcadas como não utilizáveis, independentemente da probabilidade que

lhes está associada[64]. Esta é a fase de teste. No gráfico da figura 23 podemos ver as ligações abandonadas a vermelho[65].

Chegamos à fase final denominada de "resolução de ambiguidades". Esta é a fase mais complexa, apesar de, em termos computacionais, ser relativamente rápida por comparação com a fase de geração. A questão resume-se do seguinte modo: dado que temos algumas ligações que são logicamente impossíveis e dado que, definido um limiar prévio, há ligações que em princípio são válidas e outras não, como fragmentar o grafo em conjuntos separados de ocorrências que correspondam cada um a uma pessoa real?

Vejamos, como, na prática, procede o programa, enunciado à medida que entram em acção, as regras que regem as decisões tomadas.

Primeira regra: *ordenar as ligações por ordem decrescente de valor, ignorando as que foram marcadas como não utilizáveis na fase de teste.*

---

64 Uma situação recorrente em que ligações fortes têm de ser abandonadas é a que surge com pai e filhos com o mesmo nome, que normalmente produzem pesos binários elevados (mesmo nome, muitas vezes a mesma residência). Aliás o mesmo problema surge muitas vezes no trabalho manual.

65 Neste exemplo, além das consequências do óbito de 1687 foram também eliminadas as relações entre ocorrências que os actos davam como pai e filho (c261-1 e c261-2, c1720.24-p2 e c1720.24-p1). Note-se que as consequências de termos um óbito entre as ocorrências sob análise são propagadas muito conservadoramente, e resumem-se à eliminação das pessoas cuja função implica necessariamente estarem vivas, como seja noivos e noivas, padrinhos e madrinhas, testemunhas de casamento e devassa. De facto os registos com que trabalhámos registam muito desigualmente se determinada pessoa referida num acto estava ou não viva à data do mesmo. Assim não consideramos uma contradição que alguém morra em determinado momento e posteriormente seja referido num acto como marido ou pai de alguém, sem referência explícita ao facto de já ter falecido.

Neste exemplo, a ligação mais forte de todas é a que une o casamento de Mateus Gomes em 1703 e o casamento em 1720, que soma 45 pontos. Este valor elevado provém da enumeração das anteriores mulheres em cada casamento, o que fornece muita informação coincidente entre os dois actos. Como esta ligação é a mais forte de todas, consideramos que as ocorrências respectivas dizem respeito à mesma pessoa. Juntamos por isso as duas ocorrências criando um primeiro agregado[66]. A ligação imediatamente a seguir, em termos de valor, liga o baptismo da filha Ana, em 1716, aos óbitos da mulher Ana Domingues e da filha Teresa em 21 de Outubro de 1710, que vai constituir um outro agregado[67]. Aqui a força elevada da ligação vem do facto de, nestes actos, Mateus Gomes aparecer com o nome "Mateus Gomes Bartolo", criando uma forma nominal rara que pontua fortemente. O processo continua do mesmo modo, utilizando as relações por ordem decrescente de importância (ver figura 24).

As sucessivas ligações de forte valor, processadas nas etapas iniciais da fase de desambiguação, tendem a criar agregados isolados de duas ocorrências. Com o avançar do processamento, contudo, as ligações consideradas deixam progressivamente de juntar pares de ocorrências "livres" em novos agragados e tendem cada vez mais a referir pares em que um ou ambos os elementos se encontram já dentro de um agregado. Se uma ligação a ser tratada une uma ocorrência livre e uma ocorrência já pertencente a um agregado, então aceitar essa ligação significa juntar um novo elemento ao agregado existente. Do

---

[66] Utilizamos "agregado" no contexto onde a literatura de língua inglesa utiliza o termo "cluster".

[67] A mulher e a filha, de 14 anos, morrem no mesmo dia, de modo que, durante o registo de dados, uma mesma identificação foi introduzia para Mateus Gomes, utilizando a técnica de assinalar o elemento "mesmo_que" no acto, tal como foi descrito no capítulo 3.3. secção 3.3.4.1.

mesmo modo, se uma ligação une duas ocorrências pertencendo cada uma a um agregado diferente, então a aceitação da ligação implica juntar os dois agregados num só.

Em resumo, à medida que o programa vai precorrendo a lista de todas as ligações por ordem descrescente de valor, efectuará, para cada ligação, uma das seguintes acções:

1) Se ambas as ocorrências estão livres: cria um novo agregado.
2) Se uma está livre e a outra não: junta a livre ao agregado da outra.
3) Se ambas pertencem a agregados diferentes: junta os agregados.

Na figura 24 estas três acções estão representadas por três côres diferentes

O aspecto crucial do sistema desenvolvido consiste no modo como a decisão de efectuar, ou não, cada uma destas acções, é tomada pelo programa. Recordemos que estabelecemos um limiar abaixo do qual não aceitamos uma ligação como válida. Recordemos ainda que foi demonstrado que, por vezes, ligações acima desse limiar são falsas e ligações abaixo dele são verdadeiras. De modo que agora vamos tentar relativizar o valor de cada ligação calculando um segundo valor que representa o efeito da acção que a ligação implica.

De facto, cada ligação que desencadeia uma das acções acima descritas tem um valor próprio, que é fruto da comparação entre as duas ocorrências nominais envolvidas. O sistema calcula um segundo valor específico à acção em causa. Esse segundo valor serve para relativizar o valor da ligação em si e corresponde à ideia de que, para além da ligação em causa, há que avaliar as consequências de aceitar o par como verdadeiro ou falso. Por outras palavras há que reavaliar a ligação no contexto do processo de identificação em curso.

Esse segundo valor é calculado por três regras muito simples:

1) se a acção é criar um novo agregado então o novo valor é igual ao valor da ligação.

2) se a acção é juntar uma ocorrência a um agregado, então o novo valor é a média dos valores das ligações da ocorrência a todos os membros do agregado.

3) se a acção é juntar dois agregados então o novo valor é a média do valor das ligações entre todos os membros dos dois agregados.

Estas três procuram valorizar a decisão a tomar tendo em linha de conta a consequências da mesma. Cada uma das três regras lida com o facto de que, por vezes, ao aceitarmos uma ligação como verdadeira, estarmos implicitamente a aceitar outras ligações.

Quando estamos a considerar uma ligação que une duas ocorrências, estando uma livre e outra incluída num agregado existente, então aceitar essa ligação implica que a ocorrência livre diz respeito à mesma pessoa que as outras ocorrências que fazem parte do agregado. Assim, ao aceitarmos a ligação em causa, estamos a aceitar também as outras ligações que unem a ocorrência livre às restantes ocorrências do agregado (ver diagrama 3.12). Esta é a regra dois, acima definida, e tem, na prática duas consequências importantes: se a ligação em causa é muito forte pode "recuperar" ligações mais fracas, abaixo do limiar de aceitabilidade, que de outro modo seriam eliminadas. No exemplo da figura agregado1 se a ligação ocor-1 — ocor-5 fôr suficientemente forte pode recuperar ligações implicitamente consideradas de probabilidade baixa. Inversamente, se a ligação em causa não fôr muito forte e as ligações implicitamente consideradas tiverem valores muito baixos então o sistema pode abandonar a ligação em consideração mesmo que o seu valor esteja acima do limiar de aceitabilidade. É deste modo que o sistema utiliza "inteligentemente" o valor das ligações e o valor do limiar de aceitação.

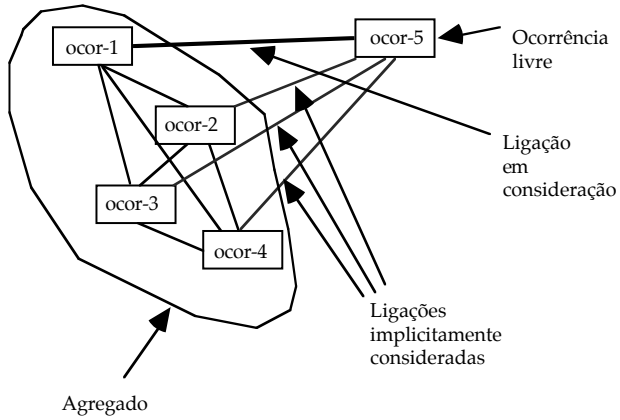


Diagrama 3.12: Ao considerar uma ligação entre uma ocorrência livre (ocor-5) e outra pertencente a um agregado (ocor-1) o sistema avalia o conjunto de ligações de facto envolvidas na decisão de aceitar ou não a ligação. Assim um novo valor para esta ligação é calculado, combinando o peso da ligação em causa com o peso das ligações implicitamente consideradas. Para a ligação ocor-1 — ocor-5 ser aceite é necessário que o valor combinado de todas as ligações entre ocor-5 e os outros membros do agregado seja superior ao limiar de aceitação.

Do mesmo modo, a consideração de um agregado implica a análise de todas as ligações entre todas as ocorrências do agregado (ver figura 3.13).

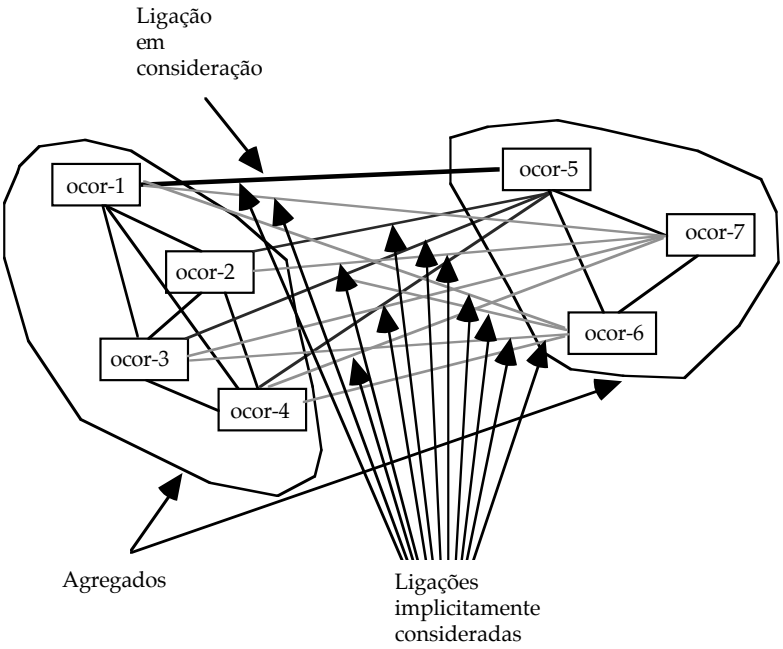


Diagrama 3.13: Quando a aceitação de um ligação implica juntar dois agregados têm de ser consideradas todas as ligações entre todos os membros dos dois conjuntos. Só se a combinação de todas essas ligações ultrapassar o limiar de aceitabilidade é que a ligação em consideração é retida.

Vejamos como estas regras são utilizadas para resolver o grafo em análise.

No canto inferior direito da figura 24 vemos as estruturas criadas a partir da referência original "c4-1". De "c4-1" a "c1717.20-p1" existe uma ligação de valor 25.8 que origina um agregado no passo 5 do processo de desambiguação. Lembremos que estas duas ocorrências correspondem ao primeiro casamento de Mateus Gomes com Maria Rodrigues, em 1695 e o segundo, com Maria Esteves em 1717. Mais tarde o sistema terá que considerar a ligação entre este segundo casamento e a ocorrência "b1719.91-p5", o baptizado em que Maria Esteves aparece como madrinha. O valor desta ligação é de 19.4 pontos, muito acima do limiar, portanto. Contudo, aceitar esta ligação implica juntar "b1719.91-p5" ao agregado "c4-1" - "c1717.20-p1" e implica, dito de outro modo, aceitar que o Mateus Gomes que casou com Maria Rodrigues em 1695 é o mesmo que é marido de Maria Esteves num baptismo em 1719. O que vale

mais? Esta ligação de 19.4 pontos ou a ligação de 8.9 pontos entre o casamento de 1695 e este baptismo, 24 anos depois?. Uma está claramente acima do limiar de 11.5 pontos, outra abaixo. Para nós é evidente que se trata da mesma pessoa, mas como formalizar esta certeza?

Aplicando a regra 2 acima enunciada iremos calcular um novo valor para a ligação. Esse valor é a média das ligações da nova ocorrência ("b1719.91-p2") aos dois membros do agregado existente ("c4-1" e "c1717.20-p1"). As duas ligações em causa têm os valores 19.4 e 8.9, sendo portanto a média igual a 14.15 pontos. Em consequência aceitamos esta ligação e recuperamos por esse facto a ligação de 8.9 pontos que de outro modo teria sido abandonada.

Inversamente poderemos ser levados a abandonar ligações acima do nosso limiar de plausibilidade através do cálculo do novo valor associado à acção. É o que acontece do lado direito da figura, com quatro ligações entre o agregado de duas pessoas à direita e o grande agregado à esquerda. Essas ligações tem um valor à volta de 13 pontos. Contudo o valor médio das ligações entre os dois agregados é extremamente baixa, situando-se à volta de 2 ou 3 pontos apenas. Sendo assim estas ligações são abandonadas.

Deste modo cada ligação é relativizada pelo contexto em que é examinada. Isto implica, do ponto de vista computacional, que o valor de todas as ligações é reavaliado a cada passo. Sempre que uma ligação é aceite altera-se o número e/ou composição dos agregados presentes e assim, as consequências das acções seguintes sujeitas a reavaliação.

A regra final que dirige a estratégia do programa é a seguinte:

A ordem pela qual as várias ligações são processadas depende não só do valor da ligação em si mas desse valor ponderado pelo valor da acção a

desencadear, calculado pelas regras anteriormente formuladas. Isto significa que haverá ligações fracas que serão inflacionadas e ligação fortes que serão deflacionadas.

O modo como a ponderação acabada de definir é feita foi alvo de determinação experimental. É um equilíbrio difícil entre o valor da ligação e o valor da acção. De certo modo a ligação exprime a realidade do facto em si, a probabilidade associada à informação documental. O valor da acção, por seu lado, exprime um saber de outro tipo, o da coerência das biografias. A nossa experiência, que poderá, ou não, ser válida para fontes com outras características, ensinou o seguinte:

1) Considerar simplesmente o valor das ligações e aceitar estas por ordem decrescente produz resultados inaceitáveis[68].

2) Substituir o valor da ligação pelo valor da acção calculado como foi explicado acima produz resultados muito superiores mas que por vezes ignoram ligações fortes que ao historiador se afiguram como inatacáveis.

3) A melhor solução parece ser um equilíbrio ponderado entre os dois valores. A média do valor da ligação e do valor da acção produz resultados ainda superiores às duas alternativas anteriores.

---

[68] Esta é a metodologia seguida pelo Grupo de Cambridge que, de certo, consequirá por outras vias manter os resultados dentro de critérios de qualidade aceitáveis. No nosso caso é indiscutível que uma estratégia de valorizar sempre as ligações mais fortes, como é defendido por Schofield em várias ocasiões, produz resultados inaceitáveis. Ver: Schofield, Roger,*Automatic Family Reconstitution*. "Historical Methods", volume 25, number 2, (Spring 1992), p. 75-79

Face a estes resultados experimentais foi decidido utilizar uma ponderaçao entre o valor da ligação e o valor da acção. Essa ponderação pode tender mais para o lado da ligação ou mais para o lado da acção através de uma parametrização do sistema. Com os dados que possuíamos afigurou-se-nos que os melhores resultados eram conseguidos fazendo uma ponderação que valorizava a ligação em detrimento da acção na proporção de dois para um.

Na prática, para cada ligação calculamos o valor da acção que a ligação desencadeia. Ficamos assim com dois valores: o valor da ligação, calculado pela soma dos pesos binários dos atributos das duas ocorrências e o valor da acção, que é calculado diferentemente conforme a ligação implicar um novo agregado, a junção de uma ocorrência a uma agregado existente ou a junção de dois agregados. Perante esses dois valores chegamos a um resultado final multiplicando o valor da ligaçao por dois, somando o valor da acção e divindo o total por três. No fundo temos a média ponderada entre os dois valores, feita de maneira que exprimimos que queremos dar mais importância à ligação do que à acção que ela desencadeia.

Como é evidente este paradigma pode ser utilizado com afinações diversas e poderemos valorizar mais a acção do que a ligação se os resultados obtidos assim o sugerirem

Esta oscilação entre valorizar o resultado do cálculo probabilístico, no nosso caso representado pelo valor da ligação, ou valorizar a consequência da acção tomada em função de cada ligação encontra-se, de forma mais ou menos explícita na literatura desde os anos 70. De facto, desde *Identifying people in the past*, que esta discussão recorre, muitas vezes personalizada como um

confronto entre os programas desenvolvidos pelo grupo de Cambridge de História Social e as propostas de Mark Skolnick[69].

Resumindo este exemplo: que novos dados se extraem deste exemplo de aplicação do sistema? Do ponto de vista da definição de um formalismo para a automatização do cruzamento nominal pensamos ter dado um ideia, esperamos que clara, de como se passa do cáculo dos pesos binários, anteriormente explicado, para a resolução concreta de um caso. Partindo de um noivo num casamento o sistema gerou automaticamente a biografia de quatro pessoas, em situações não muito fáceis: dois individuos recasam e um deles quatro vezes; o leitor atento dos documentos fornecidos em apêndice detectará um número considerável de variações nominais e de outro tipo.

A informação automaticamente fornecida pelo programa reconstitui as biografias individuais das pessoas envolvidas. Mateus Gomes, em Soure, na transição do século XVII para o século XVIII houve assim, pelo menos, quatro. Um deles, de Porto Panelas, enviuvou em 1679. A sua mulher, Marta Rodrigues foi sepultada na Igreja de Nossa Senhora de Finisterra, junto ao Castelo de Soure. Mateus sobreviveu-lhe 8 anos e, por razões que desconhecemos, quis ser enterrado nova igreja paroquial de Santiago, que D.Manuel mandara fazer, tendo dado um cruzado pela cova.

Um outro Mateus Gomes vivia não muito distante, do outro lado do rio, no Paleão. Era filho do ferreiro Manuel Gomes e casou em 13 de Outubro de 1695 com Maria Rodrigues, da Carvalheira, orfã de pai. Os irmãos de Maria, Manuel e Diogo foram testemunhas do acto. Maria não viveu muito mais de um ano, vindo a morrer em 22 de Dezembro de 1696. Embora nada o indique nas fontes,

---

[69] Ver uma comparação das duas metodologias em: Bouchard, Gérard - *Current issues and new prospects for computerized record linkage in the province of Québec*. "Historical Methods" vol. 25, nº 2 (Spring 1992), p. 67-73

o período de tempo decorrido entre o casamento e a morte de Maria aponta para uma provável complicação de parto. Mateus Gomes esperará 21 anos até se casar novamente. Em 1717 casar-se-á com Maria Esteves, do Paleão. A última notícia que temos dele é a referência que lhe é feita no baptizado de José, filho de Manuel Gomes Ferreiro e de Isabel Tomé, provavelmente seu sobrinho.

O terceiro Mateus Gomes morava na vila. Aparece por vezes com o nome de Mateus Gomes Carreiro ou ainda Mateus Gomes Bártolo. Deveria ser irmão da Misericórdia uma vez que é na igreja da Santa Casa que se casa com Maria Ferreira, orfã de pai e mãe em 12 de Novembro de 1694. Não era o seu primeiro casamento, nem será, longe disso, o último. Menos de um ano antes, em 23 de Janeiro, tinha ido a sepultar a sua primeira mulher, Maria Luís. Não temos o registo deste primeiro casamento. Mateus Gomes esteve viúvo apenas 8 meses. Não há notícia de filhos deste segundo casamento, como não havia do primeiro. Em 1703 Maria Ferreira, a segunda mulher, morre "apressadamente" segundo nota do padre Luís Alvares Pinto, o que significa que não houve tempo para lhe ministrar os últimos sacramentos. Estamos em 19 de Outubro de 1703. Menos de dois mêses depois Mateus Gomes casa de novo com Antónia Domingues, em 16 de Dezembro de 1703. Os filhos aparecem enfim: Teresa em 1704, Catarina em 1707, e Martinho em 1709. Catarina morre com 3 anos de idade em 1710. Nova filha em 1712, Maria, morrerá igualmente três anos depois em 1715 e quase a seguir morre Martinho em 18 de Outubro desse ano. com seis anos de idade. Em 1716 nasce Caetana. Mas foi em 1718 que a tragédia se abatou sobre Mateus Gomes: a mulher Antónia Domingues e a filha primogénita Teresa morrem no mesmo dia. Assim Mateus Gomes fica de nosso viúvo. Dos seus cinco filhos só sobrevive Caetana com pouco mais de um ano e que morre em 1719 com dois anos de idade. Assim Mateus Gomes recomeça de novo e casa com Isabel Rodrigues em 1720. A partir daqui faltam-nos os registos para seguir o seu rasto.