DO MANUSCRITO AO TECLADO

Os usos da informática na investigação histórica

Nuno Camarinhas

1. ANTECEDENTES

Desde meados dos anos 70 que se verificam tentativas de introduzir o recurso ao computador e à informática como ferramentas de análise a ter em conta na recolha e, sobretudo, no tratamento de dados de carácter histórico. Num período em que ainda estavam em voga os megalómanos projectos de história quantitativa (dita serial), esse recurso era grandemente dificultado pela raridade dos meios informáticos disponíveis. "Computador" era uma palavra que se usava no singular dados os valores elevadíssimos que este tipo de máquina podia custar.

Quando a micro-informática se começa a desenvolver e o computador passa a estar cada vez mais ao alcance do utilizador particular, a historiografia de carácter mais quantitativo conhece, por seu lado, uma acentuada recessão. São os tempos da micro-história e da história das mentalidades. São, dito de outro modo, os tempos em que as abordagens historiográficas tendem a centrar-se em objectos mais reduzidos, em estudos de caso, em suma, numa escala micro, onde o recurso a grandes cálculos é residual e, por isso, se continua a adiar o encontro assumido entre história e informática.

Este encontro, não sendo assumido, isto é, não estando generalizado, começa, no entanto, a esboçar-se. Os finais dos anos 80 e toda a década de 90 assistem ao nascimento, em diversas universidades e unidades de investigação espalhadas pela Europa e pelos Estados Unidos, de grupos de trabalho que se caracterizam por recorrerem à informática na sua investigação. Em 1987, por exemplo, foi fundada a Association for History and Computing[1] com o intuito de promover e desenvolver o

[1] A Association for History and Computing (AHC) tem um site com o endereço http://odur.let.rug.nl/ahc onde se poderá encontrar mais informação, embora actualmente (Abril de 2005) algumas das ligações não estejam a funcionar correctamente. A revista publicada pela secção britânica da associação, *History and Computing*, é particularmente interessante pelos artigos que dedica a esta aliança.

interesse pelo uso de computadores não só na investigação histórica como também na divulgação do saber histórico.

Esta aproximação recebe um impulso importante dado pelos grandes projectos de investigação em torno das buscas das origens do Estado Moderno[2]. Este projecto, financiado pela Fondation Européenne de la Science e de enorme fôlego, vai recorrer pela primeira vez de uma forma sistemática à abordagem de carácter prosopográfico, e, consequentemente, vai fazer uso da informática para a recolha e tratamento dos milhares de dados que vão ser recolhidos em torno das administrações centrais medieval e moderna. Da preparação dessas iniciativas, materializada em encontros e mesas redondas sobre os usos que os historiadores podem fazer da informática, ficou-nos importante bibliografia de carácter teórico mas igualmente prático[3].

Em finais dos anos 90 e nos primeiros anos do novo século três factores vão, finalmente, permitir o anunciado encontro entre a investigação histórica e a informática:

a) o crescimento exponencial do parque informático, graças à acentuada baixa do seu preço;

b) a cada vez maior facilidade de utilização do *software* disponível;

c) o aumento do recurso à internet.

Se em inícios dos anos 90, era bizarro entrar com um computador numa sala de biblioteca ou de arquivo, hoje em dia é estranho entrar nelas apenas com lápis e folha de rascunho. Por outro lado, os programas básicos de trabalho que por norma são utilizados atingiram um grau de sofisticação e de facilidade de uso que permitem mesmo ao utilizador menos experiente proceder a operações de razoável complexidade facilmente e com espantosa velocidade. Se nos anos 70, 80 e mesmo inícios dos anos 90 era necessário, muitas vezes, recorrer a programas desenvolvidos especificamente por especialistas em programação para cada caso, hoje os pacotes generalistas do tipo Microsoft Office são capazes de efectuar grande parte das tarefas que se esperam de um programa de computador. A ligação em rede, por fim, permite uma muito maior troca de informação não só de teor mais avançado – isto é, de resultados de investigação –

---

2 Informação mais detalhada, incluindo alguma da bibliografia resultante das várias equipas que compunham o projecto, pode ser encontrada em http://lamop.univ-paris1.fr/W3/lamop10.html.

3 Destacam-se as obras: F. AUTRAND (ed.), *Prosopographie et genèse de l'État moderne [Texte imprimé] : actes de la table ronde, Paris, 22-23 octobre 1984*, Paris, ENSJF, 1986; J.-P. GENET e G. LOTTES (eds.), *L'État moderne et les Élites XIIIe-XVIIIe siècles. Apports limites de la méthode prosopographiques. Actes du Colloque International CNRS-Paris I, 16-19 octobre 1991*, 1996; e H. MILLET (ed.), *Informatique et prosopographie*, Paris : CNRS, 1985.

mas também de carácter mais simples – manuais de utilização de ferramentas, introdução a metodologias, exemplos práticos de aplicação de determinadas técnicas, tudo ficou, num espaço muito curto de tempo, disponível para o utilizador interessado.

2. ALGUMAS APLICAÇÕES

As capacidades actuais da micro-informática possibilitam um acompanhamento muito próximo das diferentes fases da investigação histórica. Vou debruçar-me sobre dois aspectos em que tenho desenvolvido mais esforços: a digitalização de texto e o tratamento de dados documentais para inserção em bases de dados.

2.1. Digitalização de imagens e textos

Em 1996, a extinta Comissão Nacional para as Comemorações dos Descobrimentos Portugueses iniciou um projecto de edição de CD-ROM's textuais do qual tive o privilégio de fazer parte[4]. Tratava-se disponibilizar ao utilizador um corpus textual de dimensões consideráveis (uma colecção inteira de uma revista científica, uma obra de transcrição de documentos em vários volumes, grandes dicionários bibliográficos, as obras completas de um autor). Às vantagens óbvias de ter toda esta informação num só CD, acresciam outras bem mais consideráveis: os CD-ROM's continham o texto integral em formato digital o que permitia fazer pesquisas por qualquer palavra ou frase. Era possível, inclusivamente, classificar a informação, e, apesar de se tratar de texto livre, não formatado à maneira das fichas de bases de dados, pesquisar dentro de campos específicos (como os topónimos, ou os títulos das obras, por exemplo). O formato digital permitia, por outro lado, acrescentar mais valias aos diferentes CD's, quer se tratasse de cronologias, textos introdutórios ou explicativos, bibliografia ou, até, bases de dados que orientavam o utilizador e permitiam formas alternativas de navegar pelo texto.

---

4 O projecto *Ophir – Biblioteca virtual dos descobrimentos portugueses*, publicou, entre 1997 e 2003, os seguintes títulos: *Stvdia (n.º 1 a 53)*, coord. de Ruth Martinho, 1997; *Bibliotheca Lusitana (Diogo Barbosa Machado)*, coord. de André Belo, 1998; *Boletim da Filmoteca Ultramarina Portuguesa (n.º 1 a 50)*, coord. de Catarina Madeira Santos, 1998; *Corografia Portuguesa (Padre António Carvalho da Costa)*, coord. de Ana Cristina Nogueira da Silva, 2002; *Décadas da Ásia (de João de Barros)*, coord. de Thomas Earle e Stephen Parkinson, 1999; *Mare Liberum (n.º 1 a 13)*, coord. de João Paulo Salvado, 1999; *Gil Vicente, Todas as Obras*, coord. de José Camões, 2002; *Dicionário Bibliográfico Português (Inocêncio F. Silva)*, coord. de André Belo, 2002; e *Historia de Japam (Fr. Luís Fróis)*, 2003. Estão, ainda hoje, por editar, os volumes relativos às *Obras Completas de Padre António Vieira*, às *Monumenta*

A opção pela digitalização integral do texto conferia todas estas vantagens aos CD-ROM's mas tinha duas grandes desvantagens: reduzia o universo dos textos utilizáveis aos impressos e tinha enormes custos nos tempos de produção. Não bastava digitalizar as páginas dos livros, isto é, obter uma imagem "fotográfica" de cada folha. Era necessário fazer o reconhecimento óptico dos caracteres (OCR)[5], processo pelo qual um programa de computador lê a fotografia da página e transcreve-a para texto electrónico. Dependendo da qualidade da página original, obter-se-iam mais ou menos erros de leitura que, depois, teriam de ser corrigidos manualmente. Outra opção possível passaria pela digitalização apenas como conjunto de imagens às quais seriam associados descritores, esses sim, pesquisáveis. Mas o que se perdia em capacidade de pesquisa levou à adopção do método descrito.

A digitalização, quer de imagens, quer de textos, é um utensílio de extrema utilidade. É uma alternativa de conservação e de difusão de documentos. Permite reunir, mesmo num computador portátil, uma colecção enorme de registos que, por sua vez, podem ser indexados, organizados e pesquisados através das mais diversas e comuns ferramentas informáticas.

2.2. As bases de dados relacionais

Outro campo de aplicação da informática à investigação histórica, talvez o campo por excelência, é a constituição de bases de dados. Uma base de dados não é mais do que uma sistematização da informação recolhida. É a versão informática do velho e tradicional arquivo de fichas de cartão, mas com uma capacidade de exploração inúmeras vezes superior.[6]

O princípio de qualquer base de dados é a decomposição de uma informação complexa em diferentes campos que a descrevem. Uma lista bibliográfica, por exemplo, pode ser uma base de dados em que os campos são o nome do autor, o título da obra, o local de edição, etc. Ou uma lista de nomeações de ofícios. Qualquer documento que tenha informação estruturada e serial é passível de ser convertido numa base de dados. Mas a informação que se pode reunir numa base de dados não tem que se limitar apenas

---

Henricina e a um conjunto de documentação goesa, intitulado *Luso-Orientalia*.

5 Software como o OmniPage o TextBridge ou o Readiris são os programas mais correntes para este tipo de função.

6 Para uma excelente introdução à criação e exploração de bases de dados sob o ponto de vista de um historiador, J. CELLIER, Traiter des données historiques, Rennes, Presses universitaires de Rennes, 2003.

a uma única fonte. Uma base de dados pode cruzar informações de fontes diversas e até de natureza diversa. A isso chama-se uma base de dados relacional. A lista bibliográfica pode ser enriquecida com informações biográficas sobre cada autor nela presente. Uma ficha de um autor fica, então, relacionada com a ficha ou fichas dos seus livros. Na primeira estão dados biográficos sobre o autor, nas segundas, dados bibliográficos sobre os seus escritos.

As possibilidades de pesquisa, de ordenação, de cruzamento de dados ou de contagem são imensas. Se no tempo das fichas em cartão, era necessário um ficheiro para cada tipo de ordenação, ocupando-se com isso várias estantes com repetições de dados, com as bases de dados informáticas, as ordenações ou as filtragens passaram a estar à distância de um clique.

Informação de origem diversa, textual mas também numérica ou cronológica, imagens ou sons, tudo pode ser classificado e incluído numa base de dados. Com os produtos correntemente disponíveis no mercado, é possível, depois, fazer migrar, com facilidade, a totalidade ou parte dos dados para diferentes tipos de aplicações, diversificando ainda mais o tipo de manipulação: as folhas de cálculo, para elaborar gráficos; os processadores de texto, para incluir uma tabela num artigo; os programas de informação geográfica, para representar num mapa uma distribuição de uma determinada variável; os programas de *social network analysis*, para traçar redes de clientela, de poder, de sociabilidade...

As bases de dados pressupõem que a informação esteja fortemente estruturada, que a cada campo corresponda sempre um determinado tipo de informação, descrita de uma forma regular. Esta rigidez parece adversa a documentos de constituição mais livre como um texto corrido. O investigador pode não querer desmontar um texto, esquartejando-o em vários campos, perdendo assim a sua unidade original. Para isso existe também, outro tipo de bases de dados, designadas de bases de dados textuais[7]. São programas mais vocacionados para um tratamento linguístico da informação. A informação é tratada sempre em contexto, embora seja possível combinar essa fluidez com algum nível de estruturação, através da introdução de marcações que, no fundo, identificam determinadas passagens do texto como pertencentes a um determinado tipo de informação (isto é, a um campo). Para além das buscas típicas de uma base de dados e de contagens básicas, este tipo de programa tem desenvolvidas ferramentas de busca

[7] Um bom programa de bases de dados textuais é o askSam.

contextualizada, de proximidade, de coexistência numa mesma frase, num mesmo parágrafo, etc. São grandes as vantagens em termos de análise textual que daqui podem ser extraídas.

O tratamento informático dos dados de carácter histórico, se traz as vantagens já descritas, não deixa de ter dificuldades, muitas delas criadas pela própria manipulação da informação. Poucas vezes, para não dizer raramente, o documento se coaduna perfeitamente com a necessidade de estruturação que a ferramenta informática exige. Cabe ao investigador a responsabilidade de não distorcer a informação para que ela se torne tratável. Muitas das dificuldades terão de ser contornadas pela introdução de campos novos que permitam uma mais perfeita tradução da informação contida no documento para um formato estruturado. Para lidar de forma mais hábil com este tipo de problemas de adaptação das ferramentas informáticas à exploração que se pretende fazer de um determinado documento os programas a que se recorre deverão ser o mais personalizáveis possível.

Raramente uma estrutura de dados criada para um propósito se pode transpor para outro objecto de estudo. Saber como estruturar é, por isso, fundamental. Grande parte do trabalho relacionado com uma base de dados se centra no desenvolvimento da sua estrutura. Para o fazer de forma correcta é preciso conhecer a fundo a documentação que se vai tratar. É sempre possível fazer alterações à estrutura, à medida que se introduzem os dados, mas há decisões que deverão ser tomadas previamente, sob risco de se perder bastante (e precioso) tempo com correcções posteriores. De uma correcta estruturação da base de dados dependerá, não apenas a fiabilidade com que a base reproduz a informação contida na documentação, mas também o potencial de exploração que ela terá.

3. CONCLUSÃO

Descrevi alguma da minha experiência no uso da informática como ferramenta de análise histórica. Foquei algumas das suas potencialidades, das armadilhas que cria e possíveis formas de as contornar. Dois pontos me parecem importantes de reter: o recurso à informática é uma forma de potenciar as capacidades de exploração de um conjunto de informação de carácter histórico. Falo da informática como utensílio de trabalho pessoal. Mas o uso de meios informáticos é, também, uma forma de divulgação do conhecimento histórico. A disponibilização de informação em formato digital, quer

através de suportes como o CD ou o DVD, quer através da rede de Internet, é outra das vertentes, talvez a que tenha mais consequências para a comunidade científica, da descoberta da tecnologia pelo investigador. Saber utilizá-la de forma a criar formatos de recolha de dados, inventários de documentação, índices de pesquisa, é a forma de possibilitar que o trabalho individual, muitas vezes isolado, se torne frutífero para os outros investigadores.