Thema: criação e gestão de *thesauri*

João Paulo Amado
João Carlos Cardoso
António Manuel Neves

Grupo HεRÓDOTO

Ficha técnica

Direcção científica e coordenação
Prof. António Manuel Hespanha

Concepção e programação
João Paulo Amado, João Carlos Cardoso, António Manuel Neves, Maria do Castelo Romeiras, Helena Isabel Medeiros

Apoio historiográfico
Paulo Artur Baptista, José Joaquim Martinheira, Pedro Cardim, Paulo Girão, Maria João Soares, Maria Alexandra Ribeiro, Ana Paula Pequito

Distribuição
Duas diskettes de 5¼" de 360Kb ou
Uma diskette de 3½" de 720Kb
(Informação compactada)

Material necessário
Computador IBM PC, PS/2 ou compatível com:
640 Kb de RAM;
Disco rígido com (pelo menos) 6Mb livres, para o programa, bases de dados e sistema de gestão de memória virtual;
Écran de qualquer standard (MDA, CGA, Hércules, EGA, MCGA, VGA, SuperVGA, XGA); para utilizar os módulos de paleografia e cartografia é necessário qualquer standard a partir do EGA;
Uma impressora para a obtenção de listagens.

Instalação
A instalação do programa é feita automaticamente, com adaptação imediata ao material disponível.

Contacto
Grupo HERÓDOTO
Apartado 1977
1006 LISBOA CODEX
PORTUGAL

Índice

1. Introdução

2. THEMA. Suas características.
2.1. Definição e objectivos
2.2. O ambiente de trabalho
2.3. A manipulação de informação
2.3.1. Edição
2.3.2. Organização de informação
2.3.3. Pesquisa
2.3.4. Capacidades genéricas
2.3.5. Impressões

3. Pormenores técnicos
3.1. A natureza do programa
3.2. Estruturação da informação e relações entre bases de dados

Conclusão

1. Introdução.

A recuperação da informação contida em documentos impressos ou electrónicos, pode beneficiar muito da aplicação de um *thesaurus*. No entanto, para que tal recuperação seja bem feita, a linguagem de indexação deverá ser correctamente elaborada.

As regras a seguir para a elaboração de *thesauri* encontram-se já bem documentadas, não só por normas ISO específicas (nº 2788 para *thesauri* monolingues e nº 5964 para *thesauri* multilingues), mas também por uma abundante bibliografia de autores nacionais e estrangeiros.

Não é nosso objectivo abordar os problemas relativos à elaboração de uma linguagem de indexação. Pretendemos, isso sim, fornecer uma ferramenta que trabalhe a um nível superior, facilitando a construção de tais linguagens.

2. THEMA. Suas características.

2.1. Definição e objectivos.

Um *thesaurus* consiste num "vocabulário controlado de termos para indexação, organizado formalmente por forma a que as relações existentes *a priori* entre conceitos sejam convenientemente explicitadas" (Aitchison, 1982). Destinam-se a ser utilizados em sistemas de recuperação de informação: bases de dados electrónicas, índices e catálogos impressos, etc.

Um *thesaurus* é útil pela limitação que impõe à quantidade de termos por que é constituído. Torna mais fácil a classificação da informação, eliminando algumas das dificuldades inerentes à utilização da linguagem natural. Por outro lado, por armazenar as relações entre termos, facilita a busca de informação de uma forma encadeada. Se houver um bom sistema de consulta de informação, torna-se fácil andar a "passear" dentro dessa mesma informação, com base nas relações existentes entre os seus elementos.

O sistema de construção e gestão de *thesauri* que aqui apresentamos começou a ser elaborado no primeiro semestre de 1991. Dadas as suas características, é lógica a sua integração desde o primeiro momento no programa **Heródoto - Estação de trabalho em História e Arquivística**. Por outro lado, existiam motivos para optar por esta integração. Com efeito, a proposta original do **Heródoto** previa a inclusão de um *thesaurus* de termos históricos. Tal ideia não foi concretizada, devido a dificuldades materiais e de disponibilidade de pessoal, sendo apenas fornecido um simples gestor de descritores. Optámos finalmente por incluir um sistema de criação e gestão de *thesauri*, módulo que se nos afigura mais flexível e de maior utilidade.

Propomos um sistema de criação e gestão de *thesauri* mono e multilingues segundo as normas ISO 2788 e ISO 5964. Através dele, as entidades responsáveis pela elaboração de tais instrumentos poderão ter o seu trabalho facilitado. Todo o trabalho intelectual de recolha e construção de termos, continua a ser da responsabilidade dos operadores humanos. O programa não é mais que um auxiliar à organização da lista e à automatização de certas tarefas.

A utilização do **Heródoto** enquanto ambiente de trabalho, permite a utilização imediata de uma série de ferramentas de gestão de informação, junto às quais pode fazer sentido a utilização de um *thesaurus*. Torna-se fácil manter sob um mesmo tecto várias bases de dados, relativas à descrição arquivística e a diversas áreas de conhecimento relacionadas com a história (paleografia, numismática, pesos e medidas, etc.). O *thesaurus* surge assim como uma ferramenta auxiliar, permitindo eventualmente "arrumar" melhor dados muito dispersos.

Enquanto simples base de dados, o **Thema** é tratado da mesma maneira que todas as outras geridas pelo Heródoto. Dispõe dos mesmos métodos gerais para organizar e recuperar informação. No entanto, assume também um carácter algo diferenciado, pelas características que assume:

- Manipula transparentemente duas bases de dados, muito embora para o utilizador isso não seja facilmente apreendido.

- A sua utilização é feita de uma forma diferenciada em termos de écran, o que o distingue de imediato das restantes bases de dados geridas pelo Heródoto.

- Dispõe de processos privados de ordenação e organização da informação, uns desencadeados automaticamente pelo programa, os outros livremente controláveis pelo utilizador.

Tudo o que diz respeito ao **Thema** encontra-se devidamente individualizado em termos de orgânica do programa. Se uma dada operação só puder ser efectuada sobre um *thesaurus*, os seus pontos de acesso só estarão activos que ele estiver a ser editado. O inverso também é válido: existem operações que não devem ser efectuadas sobre um *thesaurus* e que não podem mesmo ser efectuadas a partir do programa.

2.2. O ambiente de trabalho.

O **Thema** partilha do ambiente de trabalho do Heródoto. Vamos aqui apresentar apenas uma breve resenha das características fundamentais deste ambiente, uma vez que existe documentação mais detalhada sobre o assunto.

O objectivo essencial do Heródoto consiste em editar bases de dados, de natureza muito diversificada:

- Bases de dados orientadas para a descrição de espécies arquivísticas.

- Bases de dados orientadas para as tarefas de gestão arquivística.

- Bases de dados sobre áreas temáticas de especial interesse para historiadores e outros investigadores (paleografia, numismática, pesos e medidas, cronologias, etc.).

- Outras bases de dados, de natureza auxiliar, mas essenciais para que todo o conjunto trabalhe de uma forma equilibrada.

A par das capacidades de edição de bases de dados, fornece também outras ferramentas mais específicas:

- Para conversão de datas entre calendários históricos.

- Para criar aplicações de cartografia histórica.

- Para desenvolver e representar graficamente as relações entre membros de

uma família, isto já no campo da demografia.

- Para possibilitar uma representação aproximada das formas de diversos tipos de letra, no campo da paleografia.

O sistema de gestão de *thesauri*, bem como a lista simples de descritores, são também bases de dados, que podem ser utilizadas em integração com os restantes módulos do programa. A todas estas capacidades, acresce ainda a possibilidade oferecida a cada utilizador de manipular as suas bases de dados, definidas fora do **Heródoto**. Num futuro próximo, será inclusivamente possível definir as estruturas de novas bases de dados a partir do próprio programa.

Metáforas utilizadas pelo programa. De cima para baixo: tabela, janela, menu e lista de selecção.

Todas estas capacidades estão à disponilibidade dos utilizadores do programa. Para tornar a sua utilização tão fácil quanto possível, recorremos a uma série de metáforas, que tentam modelar elementos do mundo real no écran do computador. Temos assim:

- O mecanismo central de edição de bases de dados é uma **tabela**, forma de representação colunar do conteúdo de cada base de dados: a cada coluna pertence um campo, em cada linha um registo. Mesmo que todos os campos não apareçam no écran ao mesmo tempo, são fornecidos meios simples ver todo o conteúdo da base de dados.

- Informação acessória ao trabalho que estiver a decorrer num dado momento, é apresentada ao utilizador por intermédio de **janelas**, pequenos quadros de texto, com uma cor diferenciada em relação ao resto do écran. Podem conter informações diversas, elementos de aviso e auxílio, etc.

- Operações acessórias ao trabalho que estiver a decorrer num dado momento, também podem ser efectuadas dentro de **janelas**: preenchimento de campos com informação acessória, por exemplo.

- A escolha das operações a efectuar é feita a partir de **menus**, listas maiores ou menores que apresentam as diversas opções disponíveis. Sob este ponto de vista, o seu objectivo é em tudo idêntico ao dos menus dos restaurantes.

- A escolha de um ou vários elementos para uma dada operação pode ser efectuada a partir de **listas de selecção**, que são conceptualmente idênticas aos menus. No entanto, em várias ocasiões, é o utilizador que acaba por definir o conteúdo dessas listas.

- A introdução de texto livre também é feita no interior de uma **janela**, estratégia com a qual se pretende representar da melhor maneira possível uma folha de papel (ou, no pior dos casos, um autocolante tipo *post-it*).

A edição de bases de dados constitui a operação central do programa. Ao todo, são geridas directamente 34 bases de dados distintas, repartidas entre as várias áreas temáticas que o programa aborda. A parte de leão cabe sem dúvida à arquivística.

Já mencionámos o facto de a gestão de um *thesaurus* ser repartida por duas bases de dados.

Esta desmultiplicação não é excepção no **Heródoto**. Dada a natureza do formato .DBF das bases de dados utilizadas pelo programa (essencialmente pseudo-relacional), existem tipos de informação que só podem ser representados quando repartidos por mais do que uma base de dados:

- No caso da arquivística, os aspectos relativos à descrição de espécies documentais são repartidos por 10 bases de dados.

- Ainda na arquivística, os apectos relativos à gestão arquivística propriamente dita repartem-se por 5 bases de dados.

- Um dos utilitários, o módulo da prosopografia, lida com 2 bases de dados principais e uma auxiliar. O módulo da cartografia (ainda não terminado), reparte os dados relativos a cada mapa por 4 bases de dados.

O que acimá de tudo importa realçar, é a extrema facilidade como toda esta informação pode ser consultada de uma forma **integrada**. Tal é conseguido não só a partir dos próprios mecanismos normais de acesso a cada base de dados, como ainda através de funções específicas, que permitem muito rapidamente reunir dados de fontes diversas. O caso do **Thema** é exemplar na medida em que o utilizador só se apercebe verdadeiramente da existência de uma segunda base de dados através de algumas mensagens apresentadas no écran para operações mais específicas. Procurou-se isolar sempre que possível o utilizador da forma de distribuir a informação. Basta-lhe saber que ela está lá e quais os meios para a obter.

2.3. A manipulação da informação.

2.3.1. Edição.

Exemplo de écran de trabalho para um *thesaurus*.

O modo de **edição** de informação é central no trabalho com o **Thema**. Através dele são introduzidos novos termos num *thesaurus*, são feitas alterações e corrigidos erros.

A écran de trabalho é diferente do que é apresentado para as restantes bases de dados do **Heródoto**. Apresenta-se dividido em duas partes:

- Na esquerda, existe uma **tabela** clássica, onde cada coluna representa um campo e cada linha um registo. Dada a exiguidade do espaço só pode ser vista uma coluna de cada vez. As restantes estão escondidas, mas podem ser facilmente consultadas.

- Na parte direita do écran existe uma **janela**, destinada à edição em tecto livre da estrutura de relações de cada termo do *thesaurus*.

Todo o trabalho com o **Thema** acaba por se desenrolar entre estes dois segmentos do écran. Os termo são adicionados na tabela e depois para cada um deles é editada, ou não, consoante o caso, a estrutura de relações respectiva.

Esta estrutura é introduzida sob a forma de texto livre, mas não é armazenada como tal. Para poupar espaço as relações indicadas são codificadas e guardadas como tal na segunda base de dados. Na recuperação da informação, seja por pesquisa de informação, seja durante o simples visionamento durante a edição, a sequência codificada é devidamente desdobrada e processada como texto. Todo este processo é feito transparentemente, com uma ou outra eventual demora em certas operações. O utilizador consulta o *thesaurus* deslocando-se pela lista de descritores que aparece na metade esquerda do écran. Para cada descritor que for sucessivamente destacado, é mostrada a estrutura de relações respectiva. Esta estrutura não existe fisicamente, mas é apenas dinamicamente criada na altura.

Na introdução de informação podem ser utilizados livremente os caracteres acentuados previstos no português escrito. Todas as operações de ordenação e indexação são capazes de os levar em conta, por forma a que descritores e suas relações se encontrem sempre bem organizados.

No que toca à metodologia de trabalho, é conveniente tratar da definição da estrutura das relações ao mesmo tempo que é feita a introdução dos termos. Concluída a introdução de cada uma dessas estruturas, o **Thema** percorre-a e faz os necessários acertos em todo o *thesaurus*: cria novos descritores com inversão automática de relações, substitui e ajusta descrições já existentes. Todo este trabalho é processado de forma transparente para o utilizador, muito embora nos casos de estruturas muito grandes possam ocorrer breves interrupções.

2.3.2. Organização da informação.

As necessidades específicas de organização de um *thesaurus* são assim resolvidas automaticamente pelo programa. Ao introduzir um descritor, são definidas as relações que ele tem com outros. Finda esta operação, são geradas as relações inversas. Por exemplo, se um termo tiver outros 10 a ele relacionados, eles serão percorridos, sendo adicionada à lista de relações de cada um uma nova relação com o termo a ser processado no momento. Se uma das relações apontar para um descritor que ainda não exista, ele será criado automaticamente e inserido no conjunto global da informação.

Outras necessidades, ligadas a índices, são também manipuladas automaticamente pelo programa. O conjunto de descritores que formam um conjunto de relações é sempre mantido ordenado alfabeticamente e por tipo de relação. Outras operações mais específicas, quer de ordenação, quer de indexação, podem ser efectuadas através do ambiente que o Heródoto proporciona.

2.3.3. Pesquisa.

A informação existente na base de dados principal (que é também a única com a qual o utilizador contacta directamente), pode ser recuperada de várias maneiras. Mais uma vez, é feita uma utilização directa de quase todas as capacidades de trabalho do **Heródoto**. Referimo-nos neste caso às possibilidades de pesquisa de informação.

Para obter descrições mais detalhadas, será necessário consultar a documentação específica sobre o **Heródoto**. Aqui vamos apenas falar ligeiramente de cada modlidade de pesquisa relevante para o **Thema**.

a) Pesquisa em texto livre.

Esta modalidade de pesquisa encara uma base de dados como se se tratasse de um texto normal ignorando as divisões internas em campos. O utilizador limita-se a indicar o texto que pretende encontrar, para o que pode utilizar os operadores lógicos E, OU e NÃO, bem como os parêntesis.

b) Pesquisa com selecção por campos.

Esta modalidade de pesquisa obriga o utilizador a respeitar a divisão interna de cada registo. Assim, é necessário indicar em que campo se quer encontrar o quê. Estão disponíveis os mesmos operadores lógicos do tipo de pesquisa anterior, assim como uma gama de operadores relacionais (**igual a, diferente de, maior que, menor que** e **contido em**).

c) Pesquisa por sintaxe dBASE.

Esta é a forma de pesquisa mais poderosa oferecida pelo programa. No entanto, exige um conhecimento mínimo da linguagem dBASE. Permite aceder aos vários campos editáveis da base de dados, utilizar todos os operadores lógicos e relacionais que a dita linguagem comporta, assim como funções específicas não só do Clipper 5.01 como do próprio programa **Heródoto**. Não é uma modalidade de pesquisa trivial.

d) Pesquisa por Número de registo.

Ao contrário de algumas das modalidades anteriores, este tipo de pesquisa ignora os campos para se dedicar aos registos da base de dados. Permite ao utilizador indicar com grande precisão que porção da base de dados deseja consultar, por exemplo, os registos 10 a 20. Dada a natureza no formato .DBF, onde o número de cada registo não é fixo, mas apenas um mero elemento de ordenação, este tipo de pesquisa não pode oferecer os mesmos resultados continuadamente.

A estas opções de pesquisa juntam-se as possibilidades normais que o **Heródoto** oferece: gravação de expressões de pesquisa e sua recuperação, constituição de uma base de dados com expressões criadas, que pode ser consultada a qualquer altura e manipulada como qualquer outra base de dados, etc.

A par destas modalidades de pesquisa, comuns às restantes bases de dados geridas pelo **Heródoto**, o **Thema** dispõe de um mecanismo particular de recuperação de informação. A partir do écran normal de edição, a tecla F6 coloca no écran uma janela com uma lista de selecção. Cada elemento dessa lista é uma das relações que o termo corrente mantém. Escolhendo uma dessas relações, acede-se ao termo respectivo, sendo imediatamente apresentada a sua própria lista de relações, de onde pode ser feita nova selecção... etc. Esta função permite facilmente andar a "passear" por todo o *thesaurus*, vendo cada termo e as suas relações, traçando toda uma cadeia de relações com um esforço mínimo.

2.3.4. Capacidades genéricas.

Tal como acontece com todas as bases de dados geridas pelo **Heródoto**, também as do **Thema** beneficiam das capacidades de gravação para disco rígido, diskettes, etc. Desta forma torna-se fácil fazer cópias de segurança, distribuir exemplares de um *thesaurus*, etc.

Da mesma maneira, as formas de eliminar o conteúdo de um *thesaurus* são as mesmas que para as restantes bases de dados. Existe, no entanto, a possibilidade de eliminar apenas as relações de um termo, em vez do termo e das suas relações. É óbvio que esta operação pode facilmente acarretar grandes ajustes e actualizações por todo o *thesaurus*, coisa que o programa faz automaticamente. Para auxiliar à eliminação da informação, pode ser feita a qualquer altura uma análise da integridade dos termos e suas relações introduzidos até uma dada altura. Tal análise permite identificar:

- Termos sem relações estabelecidas.

- Termos com relações duplicadas.

- Termos com relações inexistentes.

Algumas destas situações raramente deverão ocorrer, dados os sistemas de segurança internos do programa. No entanto, este tipo de análise pode ser efectuado em qualquer altura.

2.3.5. Impressões.

Tal como se encontra a trabalhar neste momento, o **Thema** permite um bom grau de interactividade na construção e gestão de um *thesaurus*. Através do écran o utilizador pode controlar com grande precisão todas estas fases, dispondo ainda de várias maneiras de consultar a informação que é armazenada. O teclado e o écran são assim um dos principais elementos para a recuperação da informação.

No entanto, muitas vezes é necessário recorrer a outras ferramentas. A existência de um *thesaurus* não pode estar circunscrita a um ou vários computadores, pois muito embora estes sejam ferramentas de trabalho excelentes neste campo, ainda não podem ser utilizados em todas as situações.

Por esta razão, o conteúdo das bases de dados geridas pelo programa também pode ser impresso. Prevemos para já dois formatos distintos de impressão, que julgamos poderem cobrir as necessidades mais básicas.

- É possível tirar listagens do conteúdo de um *thesaurus* no seu estado bruto, que surgem como simples listas de temas com o conjunto de relações a eles associados. Dado que o mecanismo de impressão é o do **Heródoto**, é possível delimitar a informação a imprimir com grande precisão.

- É também possível tirar listagens alfabetizadas do conteúdo de um *thesurus*, no formato de apresentação mais simples previsto pela norma ISO 2788. Aqui os temas são listados alfabeticamente, como num *outline*, aparecendo as suas relações de uma forma graficamente subordinada.

Não colocamos de parte a possibilidade de criar outros formatos de impressão, que correspondam a necessidades mais específicas.

3. Pormenores técnicos.

3.1. A natureza do programa.

Dada a sua integração no **Heródoto**, o **Thema** também foi elaborado a partir da linguagem de programação Clipper, versão 5.01. As bases de dados têm o formato .DBF do programa dBASE, o que lhes assegura desde logo compatibilidade absoluta com centenas dos mais variados programa existentes no mercado.

As mesmas restrições e problemas apontados ao Heródoto, no que toca à rapidez de processamento e à gestão da memória, também se aplicam ao **Thema**, visto que um não pode ser separado do outro. No entanto, a versão 5.01 minimiza muitos destes problemas. Para mais informações deverá ser consultada a documentação específica sobre o Heródoto.

3.2. Estruturação da informação e relações entre bases de dados.

A informação relativa a cada *thesaurus* é repartida por duas bases de dados. Se lhe for atribuido o nome de **EXEMPLO** teremos então:

EXEMPLO.DBF	Base de dados onde são armazenados os descritores.
EXEMPLO.REL	Base de dados onde são armazenadas as relações de cada descritor.
EXEMPLO.DBT	Ficheiro apenso à base de dados EXEMPLO.DBF, no qual são armazenadas as *scope notes*.
EXEMPLO.NTX	Ficheiro de índice da base de dados principal EXEMPLO.DBF, que serve para ordenar alfabeticamente os termos, levando em conta os pormenores da língua portuguesa.
EXEMPLO.IRL	Ficheiro de índice da base de dados das relações EXEMPLO.REL, que permite minimizar os tempos de acesso para a recuperação da informação.

Toda a gestão das duas bases de dados e dos seus respectivos índices é feita automaticamente. Ocasionalmente, e dependendo da operação, podem-se registar tempos mortos no trabalho com o programa, dada a necessidade de actualizar informação. Nesses casos, o utilizador é devidamente avisado sobre o que se está a passar.

Conclusão.

Com o **Thema** nós pretendemos facilitar a vida a todos aqueles que lidam com a informação. Quem utilizar o **Heródoto** passa a dispor de um meio poderoso para construir linguagens documentais. Mesmo que não utilize as capacidades normais do **Heródoto**, dispões de um módulo de trabalho bastante útil.

O formato adoptado para as bases de dados deverá facilitar a comunicação com todo o tipo de programas. Segundo este ponto de vista, o **Thema** está muito longe de constituir um sistema fechado sobre si mesmo.

Esperamos que, como ferramenta, ele seja útil. Mais uma vez carregamos na mesma tecla: por muito bonito e muito prático que seja, não é ele que tem o trabalho do constituir o *thesaurus*. Ele apenas surge com um auxiliar moderadamente inteligente, que ajuda a manter a casa em ordem a a atar as pontas soltas. Tudo o mais assenta nos ombros de quem o utilizar.