

UMA BASE DE DADOS COM REFERÊNCIAS BIBLIOGRÁFICAS

Tânia Mara Guedes BOTELHO, Cientista da Informação, Serviço Federal de Processamento de Dados (SERPRO), Brasília, DF.

Plínio de ALMEIDA, Analista Suporte de Sistemas, Serviço Federal de Processamento de Dados (SERPRO), Rio de Janeiro, RJ.

Experiência realizada no Serviço Federal de Processamento de Dados-SERPRO, em relação à área de sistema de recuperação de documentos. Apresenta a metodologia de montagem de uma base de dados, formada a partir dos arquivos do sistema KWIC/KWOC que arrola referências bibliográficas nas áreas de administração, recursos humanos, economia e processamento de dados.

1. INTRODUÇÃO

Este artigo tem por objetivo relatar uma experiência realizada no SERPRO na área de sistema de recuperação de documentos. A abordagem deste trabalho é voltada para uma metodologia de montagem de uma base de dados com referências bibliográficas e trata-se de uma proposta para implementação do sistema. Esta montagem foi feita a partir da existência dos seguintes recursos:

- ô um sistema de informações com acesso por terminais que engloba um banco de dados e uma linguagem simples e conversacional para recuperação;
- ô arquivo de dados do sistema KWIC/KWOC com referências bibliográficas;
- ô geração de arquivo com assuntos do ACC (Automation Classification Code) do National Computing Centre e da CDU (Classificação Decimal Universal).

O teste elaborado para demonstração da viabilidade técnica de um sistema de recuperação de documentos foi realizado no SERPRO, utilizando-se um sistema de informações com acesso por terminais já implantado desde 1979.

A base de dados foi formada a partir dos arquivos do sistema KWIC/KWOC e arrola referências bibliográficas nas áreas de administração, economia, recursos humanos e processamento de dados. Foram arrolados no sistema aproximadamente 6000 itens, o que foi suficiente para a demonstração realizada.

2. SISTEMA DE RECUPERAÇÃO

2.1 Modelo conceitual do sistema

O sistema de informações com acesso por terminais é um sistema de armazenamento, recuperação e manipulação de informações a partir de dados genéricos oriundos das mais diversas fontes. O sistema é composto de duas entidades: a) uma base de dados; b) uma linguagem de recuperação (LPA).

2.1.1 A Base de Dados

A base de dados é composta de dois tipos de arquivos: Dicionário e Aplicações.

O *dicionário* contém conceitos relativos a um determinado assunto. Um elemento do dicionário define cada um dos conceitos e suas relações, segundo uma estrutura porventura existente para o assunto. Um exemplo do dicionário é um thesaurus de assuntos econômicos. A função do dicionário é orientar o acesso aos dados propriamente ditos criando para estes uma classificação mais precisa.

Uma *aplicação* é um arquivo que contém dados do mesmo formato sobre um ou mais assuntos. Um elemento de aplicação é composto de ó atributos, dados e descritores.

Os atributos titulam os dados e os descritores permitem que estes dados sejam acessados. Os dados propriamente ditos podem ser numéricos ou sob a forma de textos.

2.1.2 A Linguagem de Recuperação

A linguagem de recuperação é o meio pelo qual o usuário tem acesso à base de dados. É uma linguagem simples e interativa que não exige do usuário familiaridade com sistemas computacionais (*)

A recuperação dos dados se faz em duas etapas:

1) a partir de uma idéia bastante genérica, o usuário pesquisa interativamente o(s) dicionário(s) até chegar ao(s) conceito(s) para o(s) qual(is) efetivamente obtém informações;

2) a partir deste(s) conceito(s), o usuário recupera os dados propriamente ditos.

Tendo recuperado o conjunto de informações procuradas, o usuário pode, então solicitar a exibição das informações na tela do terminal ou manipulá-la.

3. IMPLEMENTAÇÃO DO SISTEMA DE RECUPERAÇÃO DE DOCUMENTOS

O sistema foi implementado a partir de três arquivos básicos:

- arquivo de entrada do sistema KWIC/KWOC com dados sobre: título do documento, autor, imprensa, localização física do documento e classificação (segundo CDU ou ACC);

- transcrição da ACC com: código e descrição;

- transcrição de algumas classes da CDU, com: código e descrição.

Definiu-se, então, uma aplicação sob o sistema de informações com acesso por terminais denominada BIB e indexada por três dicionários: dicionário de títulos, dicionário do ACC e dicionário da CDU.

O dicionário de títulos foi formatado a partir do arquivo de entrada do sistema KWIC/KWOC e todos os três dicionários passaram por um processo de indexação automática para permitir que todas as palavras da descrição fossem vocábulos pertinentes dos conceitos.

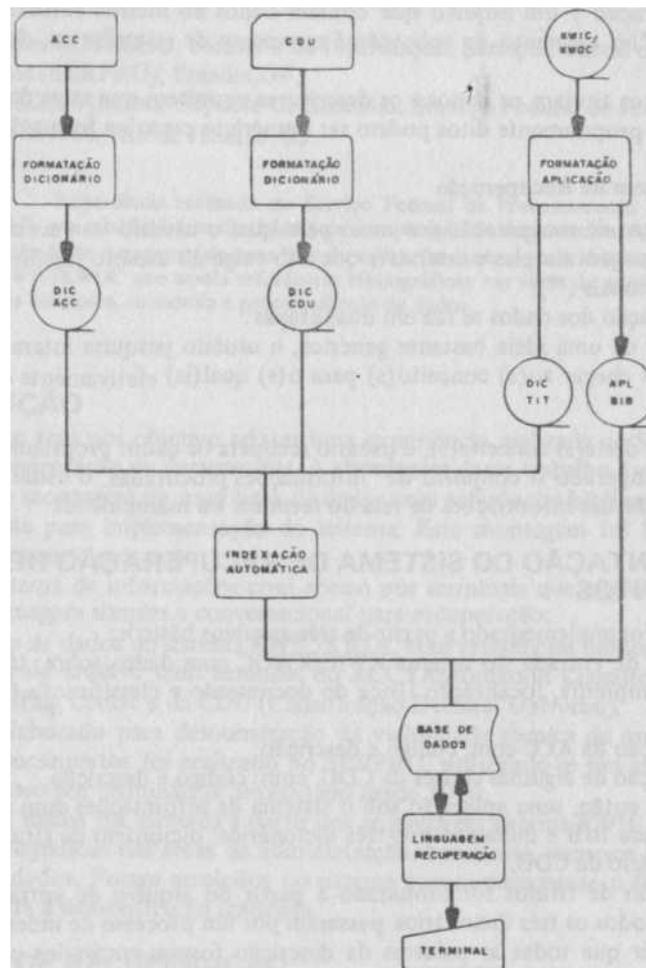
A apresentação das informações recuperadas pode ser feita em quatro formatos:

- formato, título, onde aparece a classificação do documento (ACC ou CDU), sua localização física e o título documento;

- formato autor, onde além das informações apresentadas no formato título, aparecem também os dados sobre a autoria do documento;

(*)Maiores detalhes sobre os comandos da linguagem encontram-se no Manual do Usuário publicado pelo SERPRO em 1979 e de circulação restrita.

- formato imprensa, onde aparecem dados do formato título e dados sobre a imprensa;
- formato completo, onde todas as informações são apresentadas.



Experience carried out at Serviço Federal de Processamento de Dados ó SERPRO, in the area of recuperation of information. It presents the methodology appliede to set up a data basis, builtted from the archives of KWIC/KWOC system which lists bibliographical references in the areas of administration, human resources, economics and data processing

4 BIBLIOGRAFIA

- (1) SERPRO. Manual do usuário. Rio de Janeiro, 1979.
- (2) SERPRO- Estudo preliminar para o projeto do sistema de avaliação da performance de recuperação de informações do Projeto Aruanda. Rio de Janeiro, 1979.
- (3) LANCASTER, F.W. *Information retrieval systems, characteristics, testing and evaluation*. New York, John Wiley, 1968.