Como Escolher os Campos para um Banco de Dados

Anna da Soledade Vieira
Professora da Escola de Biblioteconomia da UFMG

RESUMO

Partindo de dados obtidos no Arquivo da FINEP *(Financiadora de Estudos e Projetos) e com a finalidade de montar um banco de dados sobre documentação típica de administração de projetos, duas metodologias de bases estatísticas são apresentadas para a definição de campos de informação. Uma, baseada na ordenação dos campos segundo sua freqüência nas consultas dos usuárias, seria cabível quando estes tivessem interesses e atividades comuns. A segunda metodologia — teste do* x^2 *— seria aplicável quando os usuários tivessem interesses e atividades diversificados.*

Não obstante alguns dos mais sérios problemas da sociedade industrial ainda estarem associados a métodos desatualizados ou inadequados de tratar e interpretar informações, o interesse emergente pela formação de coleções de documentos relevantes ligadas a serviços de informação nas áreas de Ciência e Tecnologia são indicativos de que o registro dos eventos naquele domínio tornou-se um requisito básico da civilização. (1)

Cada elemento isolado que o sistema registra constitui um dado, o qual, ao ser elaborado ou conjugado a outros para fins de comunicação, transforma-se em informação. Poder-se-ia dizer que, em um sistema de recuperação de informações o dado é a matéria prima e a informação é o produto capaz de gerar uma ação, modificar um comportamento ou propiciar uma tomada de decisão.

No processamento e transferência da informação tendo em vista a pesquisa cientifica, o desenvolvimento econômico e o bem-estar social, arquivos, bibliotecas e centros de documentação tem igual responsabilidade *e* importância.

Entretanto, estas três instituições não tem recebido idêntico tratamento, seja em âmbito oficial, seja no setor privado, ficando os arquivos relegados a um plano inferior.

Com relação ao Brasil, J. Esposel (2) denuncia o descaso no tratamento da documentação administrativa e história nacional, embora esforços isolados venham recentemente sendo feitos no sentido de modernização do tratamento de arquivos oficiais e empresariais.

Por se constituírem de documentos gráficos, geralmente não publicados ou de publicação limitada, os arquivos representam uma importante parcela dos registros nacionais, seja no aspecto da documentação histórica, econômica e social, seja no que concerne à documentação cientifica *e* técnica.

Do ponto de vista empresarial, os arquivos são também imprescindíveis uma vez que as possibilidades de êxito se baseiam na programação do trabalho, na precisão das informações e na rapidez com que estas são fornecidas.

Com este enfoque, a Financiadora de Estudos e Projetos — FINEP decidiu-se pela reorganização de seu Arquivo e a criação de um banco de dados, que informasse sobre a documentação ali arquivada e sobre os projetos financiados pela Empresa.

Planejado o novo sistema e implantado um primeiro módulo, como projeto piloto, surgiu a necessidade de avaliação de sua estrutura e de seus produtos. Esse trabalho foi a motivação inicial para a pesquisa a seguir relatada, a qual foi anteriormente objeto de tese (3) para obtenção do Grau de Mestre pelo Curso de Pós-Graduação em Ciência da Informação, do Instituto Brasileiro de Bibliografia e Documentação.

Banco de dados, no presente trabalho, é tomado como uma coleção de informações inter-relacionadas de maneira coerente e que podem ser recuperadas sob quaisquer chaves de classificação ou condições lógicas, embora armazenadas de maneira não redundante. (4) Um banco de dados é constituído de unidades físicas denominadas arquivo. Estes, por sua vez, são formados por um conjunto de registros lógicos, os quais se constituem de campos, isto é, áreas do banco de dados, destinadas a receber dados com idênticas características.

AMBIENTE DO SISTEMA

A fim de que quaisquer generalizações aqui feitas para arquivos de administração de projetos possam ser compreendidas dentro de suas limitações, é necessário que se descreva a FINEP e seu Arquivo, ambiente em que se desenvolveu a pesquisa e a partir do qual todas as conclusões foram extraídas.

A FINEP, órgão da Secretaria de Planejamento

da Presidência da República, é constituída por treze setores, a saber: Presidência (PRES), Více-Presidência (VICE-PRES), Departamento jurídico (DEJ), Departamento Administrativo (DAD), Departamento Financeiro-Contábil (DFC), Centro de Processamento de Dados (CPD), Grupo de Ciência e Tecnologia (GCT), Grupo de Estudos e Projetos (GEP), Grupo de Pesquisa (GP), Grupo de Estudos de Fontes Alternativas de Energia (GE), Núcleo do Banco Interamericano de Desenvolvimento (N.BID), Grupo de Assessoria para o Gás Combustível (G. GAS) e Programa Nacional de Treinamento de Executivos (PNTE). Seu campo básico de atuação é o financiamento de estudos, projetos e programas de desenvolvimento econômico, social, científico e tecnológico, apresentados por entidades públicas e privadas. Embora cada programa tenha características próprias, em geral todos os projetos incluem documentos de natureza administrativa, financeiro-contábil e jurídica. Toda a documentação gerada desde a solicitação inicial e durante todas as etapas da vigência dos contratos vai agregar-se ao Arquivo, de uso reservado aos funcionários da Empresa.

Para otimizar as tarefas de administração de contratos foi criado um banco de dados, compreendendo quatro partes ou arquivos: um, referente aos eventos esperados e ocorridos sobre cada contrato (ADM-CONT); o segundo, relativo ao controle financeiro-contábil (CASH-FLOW); o terceiro, contendo as características de todos os projetos em andamento (CADASTRO)) e, finalmente, o quarto, objeto desse estudo, referente à documentação do Arquivo (DOCUMENTOS).

TEMA DA PESQUISA

Duas perguntas básicas deveriam ser respondidas durante a avaliação do banco de dados da FINEP, nos aspectos concernentes a seu arquivo

DOCUMENTOS:

— os campos incluídos eram realmente os devidos?
— qual a força de recuperação seletiva desses campos?

A primeira etapa dessa investigação foi a busca de uma metodologia para a definição do conjunto ideal de campos para o banco de dados da FINEP, segundo as necessidades do ambiente aqui descrito. O presente trabalho descreve essa pesquisa e, assim sendo, enfoca:

— arquivos especializados em administração de projetos;
— definição de campos de informação para um banco de dados em computador.

OBJETIVO DA PESQUISA

A investigação se constitui em uma tentativa de desenvolvimento de metodologias alternativas, com vistas a estabelecer um modelo de sistema de recuperação de informações para arquivos de administração de projetos, podendo, entretanto, as metodologias resultantes servir de orientação para arquivos de outras áreas.

No atual contexto, sistema de recuperação de informações deve ser compreendido como o conjunto ideal de campos que comporão o banco de dados e aos quais a indexação deverá se estender para a caracterização exata de cada documento. Desde que o propósito de qualquer sistema de informação é prover o usuário com documentos relevantes ao seu interesse, ele deve ser solicitado a estabelecer os parâmetros da recuperação e *a* julgar o produto recebido. As medidas mais comumente usadas para avaliar a relevância da recuperação são precisão e revocação ("recall"). Precisão refer-se à capacidade do sistema de rejeitar os documentos não-relevantes à pergunta, enquanto revocação mede sua capacidade de recuperar todos os documentos relevantes (5).

A observação do comportamento dos usuários do Arquivo da FINEP leva à conclusão de que a precisão é mais importante que a revocação na recuperação de documentos para efeitos de administração de projetos. Assim é que, para o administrador saber se uma atitude do mutuário apoia-se nos termos do acordo firmado, somente através do contrato referente àquele projeto específico poderá ser esclarecida sua dúvida; nenhum outro documento do mesmo projeto ou qualquer contrato de outro projeto dar-lhe-á as informações necessárias. Solicitações de todos os documentos de um certo conjunto são pouco freqüentes, o que confirma a menor importância da revocação para o sistema.

Partindo-se da premissa anterior e em se tratando de documentos com caracteres diferenciais (facetas) muito numerosos e diversificados, é pressuposto que a maior precisão está diretamente relacionada com três fatores:

— a exaustividade na definição do banco de dados, isto é, a criação de tal variedade de campos de informação que possibilite exaustividade na indexação e na estratégia de busca;
— a profundidade da indexação, isto é, cada documento deverá ser descrito sob todas as suas facetas (exaustividade) e, dentro de cada faceta, da maneira mais exata (especificidade). Este cuidado levará ao equilíbrio ideal entre urna força generalizadora *e* outra restritiva, atingindo-se aquele nível ótimo de indexação que, segundo Cleverdon (6), existe para cada sistema;

— a exatidão da estratégia de busca, ou seja: exaustividade quanto ao número de características do documento, especificidade quanto ao nível dentro de cada uma das características e lógica no estabelecimento das conexões entre os termos. Esta afirmativa confirma as conclusões a que chegou Lancaster (7) : para recuperação de informações são importantes tanto o alto nível de exaustividade quanto o de especificidade na busca, uma vez que eles reduzem a classe dos documentos aceitáveis, conduzindo à alta precisão e à baixa revocação pois quanto menor o número de documentos recuperados, maior a probabilidade de precisão.

Considerando-se que, em banco de dados, as funções de indexação e de recuperação são dependentes da existência do campo no sistema, concluiu-se que a definição dos campos de maneira exaustiva *é* condição necessária, embora não suficiente, para uma recuperação precisa.
Pode-se deduzir, então, que o modelo capaz de atender à precisão requerida pelos usuários de arquivos de administração de projetos ou de arquivos de quaisquer outras áreas com idêntica necessidade de precisão seria um sistema que abrigasse todos os campos de informação existentes nos documentos arquivados, a fim de permitir, nas fases posteriores do trabalho, a perfeita identificação de cada documento, sob todas as suas facetas, tanto em relação aos seus aspectos formais quanto aos de conteúdo e situacionais. Cada uma das características rejeitadas na definição do sistema redundaria em indexação e busca deficientes, resultando, portanto, em recuperação com baixa precisão.
Pesquisa interna do Centro de Processamento de Dados (CPD) da FINEP concluiu pela avaliação do custo de cada novo termo (unidades de informação armazenadas em um campo) de seu banco de dados em cerca de Cr$ 0,19. Uma vez que a cada novo campo incluído no sistema corresponde um aumento de custo igual ao produto de Cr$ 0,19 pelo total de seus termos a serem indexados, o equilíbrio na definição do banco de dados deverá ser encontrado através do grau de utilização dos campos pelos usuários.
Essa decisão apresenta uma novidade em relação às metodologias descritas na literatura de Ciência da Informação. Enquanto usualmente se considera a freqüência de termos em documentos, no presente trabalho propõe-se investigar a partir da ocorrência dos campos nas perguntas dos usuários.

MATERIAL

Com a finalidaed de identificar as necessidades dos usuários do Arquivo, bem como seu comportamento em relação à busca de documentos, foram coletadas todas as 224 perguntas feitas por eles ao Arquivo, tanto pessoalmente quanto por telefone, durante o mês de julho de 1974.
Sendo a análise e a execução dos projetos atividades continuadas, cada mês se iniciam e se concluem contratos. Não há, portanto, épocas de pique, nem de baixa procura ao Arquivo. Daí se justificar uma amostra aleatória simples, tendo sido escolhido o mês de julho, após consulta à tabela de números equiprováveis de Hald (8).
Os pedidos, no total de 224, foram anotados exatamente conforme o solicitante se expressou. A seguir cada pergunta foi analisada para identificação dos campos que a compunham e registrada a ocorrência desses campos nas perguntas.
Paralelamente, foi feita a análise da documentação fornecida, identificados os campos de informação existentes e verificada a freqüência de sua ocorrência nos 224 documentos.
Os cálculos foram parcialmente executados em computador IBM/360-40, do Centro de Computação da UFMG, utilizando-se o Programa de Tabulação Cruzada - PRTC.
Os campos identificados nas perguntas foram reconhecidos também na documentação, diferindo apenas na freqüência. São ao todo 29, a saber:

— tipo de documento: o aspecto formal do documento. Exemplo: carta, contrato, ofício etc.

— veículo da informação: o canal de comunicação. Exemplo: Diário Oficial da União onde se publicam os contratos aprovados;

— número do documento: número com que a instituição de origem caracteriza o documento;

— número do protocolo: número através do qual a FINEP incorpora o documento ao seu acervo;

—' data do documento: data de origem;

— data do protocolo: data da incorporação do documento ao Arquivo da FINEP;

— instituição de origem: nome da entidade da qual provém o documento;

— instituição de destino: nome da entidade à qual o documento se destina;

— signatário: nome da pessoa que assina o documento;

— pessoa destinatária: nome da pessoa a quem o documento é endereçado;

— cargo do signatário: posto ocupado pelo signatário do documento;

— cargo do destinatário: posto ocupado pelo destinatário do documento;

— assunto: as ações administrativas com que o documento se relaciona ou o campo do conhecimento sobre o qual versa;

— referências: correlação de conteúdo entre documentos, um mencionando outro;

— anexos: correlação física entre documentos, um apenso a outro;

— código do projeto: código alfanumérico representativo do projeto;

— nome do projeto: nome oficial do projeto;

— variações do nome do projeto: apelidos que o projeto recebe internamente;

— mutuário: entidade responsável pelo projeto;

— executor: setor subordinado ao mutuário, onde é implantado o projeto;

— setor FINEP: nome do programa da FINEP ao qual o projeto está vinculado;

— classificação do projeto: área em que o projeto se enquadra, seja na classificação interna do GEP, seja na do Plano Básico de Desenvolvimento Científico e Tecnológico (PBDCT) utilizada pelo GCT;

— fonte de recursos: Instituições nacionais e internacionais de onde se provêm os recursos aplicados ao projeto;

— Estado(s) do Brasil: unidade federativa onde se realiza o projeto;

— agente financeiro (AF): Banco de Desenvolvimento regional responsável pelo repasse de verbas;

— agência do AF: subdivisão estadual dos Bancos regionais;

— nome da consultoria: nome do escritório técnico que dá consultoria ao projeto;

— registro da consultora: número de registro que a consultora tem no cadastro da FINEP;

— valor: quantia a que o documento se refere. Exemplos: valor do financiamento, no contrato; ou valor pago, em um recibo.

TRATAMENTO, ANALISE E INTERPRETAÇÃO

Os dados da amostra foram tratados estatisticamente, buscando elementos que permitissem a identificação de quais os campos ideais que o banco de dados deveria incluir, a fim de cobrir todas as facetas da documentação útil. Os passos seguidos para cumprimento do objetivo foram:

a) comparação entre o potencial informativo existente nos documentos e seu uso efetivo pelos usuários;

b) análise das perguntas para verificação da existência de um núcleo de campos, comum a todos os Departamentos.

Duas metodologias básicas foram seguidas utilizando instrumentos estatísticos para tratamento dos dados obtidos com as perguntas dos usuários:

a) estudo da freqüência relativa dos campos nas perguntas, sem distinção de Departamento ou tipo de documento;

b) teste do x^2 (qui-quadrado) para identificação de necessidades comuns a todos os Departamentos com referência aos campos de informação.

POTENCIALIDADE VERSUS USO DOS CAMPOS

A observação das Tabelas 1-2 permite a análise comparativa da ocorrência dos campos de informação nos documentos e nas perguntas.

TABELA 1: CAMPOS IDENTIFICADOS NAS PERGUNTAS DOS USUÁRIOS DO ARQUIVO (FINEP, RIO DE JANEIRO, JULHO DE 1974)

Nome dos Campos	f	fr	frp
Tipo de documento	148	0196	0661
	119	0158	0532
Assunto	72	0095	0322
	70	0,093	0313
Instituição de origem	54	0072	0242
	41	0054	0184
	38	0050	0170
Nome da consultora	27	0,036	0 121
Agente financeiro	25	0,033	0,112
Instituição de destino	21	0,028	0094
Variações do nome do projeto	16	0,021	0,071
Pessoa destinatária	14	0,019	0063
Número do protocolo	14	0,019	0063
Veículo da informação	11	0015	0049
Nº registro da Consultora	11	0,015	0049
Nome do projeto	10	0,013	0045
Data do protocolo	9	0,012	0040
Valor	9	0,012	0040
	8	0,011	0036
Signatário	7	0009	0031
Anexos	7	0009	0031
Cargo do signatário	5	0,007	0,022
Cargo do destinatário	5	0007	0022
Executor do projeto	4	0,005	0,018
Setor FINEP	3	0004	0013
Agência do AF	3	0,004	0,013
listados do Brasil	2	0,003	0,009
Classificação do projeto	1	0,001	0,004
Fontes de recursos	1	0,001	0,004
T O T A L	755	1,000	

Fonte: Pesquisa da autora no Arquivo da FINEP.

* O somatório dessa coluna não é significativo,

fr : freqüência relativa ao somatório de f (755)

frp : freqüência relativa ao total de perguntas feitas ao Arquivo (224)

TABELA 2: CAMPOS IDENTIFICADOS NOS DOCUMENTOS DO ARQUIVO (FINEP, RIO DE JANEIRO, JULHO DE 1974)

Nome do Campo	f	fr	frp
Tipo de documento	224	0,055	1,000
	224	0,055	1,000
Instituição de origem	220	0,054	0,982
Setor FINEP	213	0,052	0,951
Nome do projeto	207	0,051	0,924
Classificação do projeto	207	0,051	0,924
	207	0,051	0,924
Data do documento	202	0,050	0,902
Signatário	202	0,050	0,902
Mutuário	202	0,050	0,902
Estados do Brasil	202	0,050	0,902
Código do projeto	200	0,049	0,893
Instituição de destino	184	0,045	0,821
Cargo do signatário	153	0,038	0,683
Pessoa destinatária	135	0,033	0,603
Número do documento	121	0,030	0,540
Variações do nome do projeto	106	0,026	0,473
Número do protocolo	97	0,024	0,433
Data do protocolo	97	0,024	0,433
Nome da consultora	97	0,024	0,433
Nº>registro da consultora	97	0,024	0,433
Cargo do destinatário	90	0,022	0,402
Referências	72	0,017	0,321
Agente financeiro	70	0,017	0,313
Anexos	59	0,014	0,263
Valor	52	0,013	0,232
Executor	50	0,012	0,223
Agência do AF	45	0,011	0,201
Veículo da informação	34	0008	0,152
T O T A L	4069	1,000	—

Fonte: Pesquisa da autora no Arquivo da FINEP

* O somatório dessa coluna não é significativo

fr : freqüência relativa ao somatório de f (4069)

frp : freqüência relativa ao total dos documentos fornecidos (224)

Verifica-se pela simples observação dessas tabelas que a ocorrência dos campos nos documentos mostra-se bem maior que a dos campos nas perguntas em relação a todos os itens da série. A comparação entre o potencial de uniformações existentes nos documentos e sua utilização pelos usuários leva a três conclusões parciais sobre a situação particular do Arquivo da FINEP:

a) apenas uma parcela desse potencial é utlizada pelos usuários quando buscam um documento no Arquivo;

b) poucos campos são muito utilizados, enquanto que muitos campos são sub-utilizados;

c) a definição do banco de dados a partir da freqüência dos campos nos documentos seria anti-econômica, uma vez que a baixa utilização de alguns campos não justifica o custo de sua manutenção no sistema. Quanto a este aspecto pode-se observar (Tabelas 1-2) a disparidade no uso de certos campos. Alguns, com baixa utilização pelos usuários, apresentam alta freqüência nos documentos. Fontes de recursos e classificação do projeto, por exemplo, foram usados uma única vez (fr = 0,001 e frp = 0,004), embora houvesse 207 possibilidades (fr = 0,051 e frp = 0,024) Isso eqüivale a dizer que esses campos, ainda que ocorrendo em 92,4% da documentação solicitada, foram explicitados em apenas 0,4% das solicitações. Fato semelhante ocorre com Estados do Brasil e setor FINEP. Já por outro lado, campos de baixa freqüência nos documentos (Tabela 2) foram utilizados com freqüência relativamente alta (Tabela 1), se considerada sua baixa disponibilidade. São exemplos: agente financeiro, veículo da informação, valor e referências.

IDENTIFICAÇÃO DOS COMPONENTES A PARTIR DAS QUESTÕES DOS USUÁRIOS

A partir dos dados originais levantados pela análise das perguntas, a ocorrência dos campos foi observada segundo distribuição geral e por Departamentos solicitantes.
Nos diferentes casos, os campos foram ordenados decrescentemente segundo sua ocorrência nas perguntas e essa ocorrência estudada pelas freqüências simples e relativa, bem como através do x^2- Desse estudo resultaram os elementos necessários para a avaliação dos dados:

a) *Distribuição Geral*

Estando a série em ordem decrescente e analisando-se esta a partir da freqüência simples (f, Tabela 1), pode-se verificar que os campos não foram igualmente preferidos pelos usuários: há uma concentração no uso de alguns campos e uma baixa procura de outros.
A freqüência relativa (fr, Tabela 1) é, então, capaz de expressar o peso ou valor relativo do campo no sistema, sendo esse valor determinado pela preferência do usuário.
Quando os campos escolhidos para o sistema forem um sub-conjunto (a) dos campos utilizados (A) pelos usuários e considerando-se que o somatório das freqüências relativas de todos os campos usados eqüivale à unidade, o desempenho do sistema poderá ser previsto segundo esse critério básico: quanto mais o somatório dos pesos dos campos incluídos (a) no sistema se aproximar de l (frA) melhor será seu desempenho, ocorrendo o inverso quando o somatório dos pesos tender a zero. Isso leva a concluir que a escolha dos campos feita a partir da freqüência deverá recair sobre os elementos mais freqüentes na utilização, isto é, aqueles que obtiveram maior peso (fr) segundo a preferência dos usuários.
Assim sendo, analisando-se a série a partir da freqüência relativa (fr, Tabela 1) constata-se que 50% da soma dos valores ou pesos (fr) abrangem apenas os 6 primeiros campos da série, e os 75% do total dos valores já correspondem a 16 campos. Os *25%* restantes do total dos valores incorporam os 13 campos de valores (fr) mais baixos da série.
Os 6 campos correspondentes aos 50% da soma dos pesos têm conceito muito extenso. São eles: tipo de documento, código do projeto, assunto, mutuário, instituição de origem e data do documento. Tais campos devem ser incluídos; do contrário, sua extensão não limitada prejudicaria o desempenho do sistema. Porém, sua inclusão, com o objetivo de se obter recuperação precisa, implica em especificidade na indexação dos termos que cada campo abrigará.
Os 10 campos seguintes na série (do sétimo ao décimo sexto item) correspondem a mais de 25% do total dos pesos. São eles: número do documento, nome da consultora, agente financeiro, instituição de destino, variações do nome do projeto, pessoa destinatária, número do protocolo, veículo da informação, número de registro da consultora, e nome do projeto. Sua posição mediana na série parece indicar que estes campos merecem uma preferência equilibrada por parte dos usuários e têm, por conseguinte, seu lugar assegurado no sistema.
Os restantes 25% do total dos pesos correspondem àqueles 13 últimos campos da série. São eles: data do protocolo, valor, referências, signatário, anexos, cargo do signatário, cargo do destinatário, executor do projeto, setor FINEP, agência do AF. Estados do Brasil, classificação do projeto, e fontes de recursos. Tais campos, pouco freqüentes nas perguntas, podem ser considerados de conteúdo informativo menor ou efêmero e, portanto, de importância secundária para o sistema.

A inclusão daqueles campos de baixa freqüência — os últimos da série, cuja soma dos pesos perfaz *25%* dos valores de todos os campos — deverá ser considerada pelo administrador do banco de dados, do ponto de vista de custo/benefício: compensa para a Empresa investir em informações pouco utilizadas? Ou *serf,* admissível uma recuperação menos precisa em favor de mais baixo custo do sistema?

b) *Distribuição por Departamento*

A Tabela 3 dá uma visão completa da utilização de cada campo pelos diversos Departamentos, bem como do volume de solicitações de cada um.

Ordenando-se os campos decrescentemente, por linha, ter-se-ia o perfil dos Departamentos, isto é, os campos preferidos por cada setor individualmente. A freqüência de utilização dos campos se apresenta bastante dispersa. Essa dispersão é demonstrada pela baixa freqüência dos campos em geral. Esse fato parece estar relacionado à diversificação de atividades e necessidades dos Departamentos. Ainda como apoio a essa suposição, pode-se verificar a semelhança existente entre os dados referentes ao GEP e ao GCT, Departamentos com atuação muito semelhante, ambos trabalhando diretamente com análise técnica e avaliação de projetos.

TABELA 3: OCORRÊNCIA DOS CAMPOS SEGUNDO OS DEPARTAMENTOS

(FINEP, RIO DE JANEIRO, JULHO DE 1974)

Departamentos \ Campos de informação			Agência do AF		Agente Financeiro		Anexos		Assunto		Cargo do destinatário		Cargo do signatário		Classificação do projeto	
Sigla	Perguntas	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr
PRES	7	0,032	0	0,000	0	0,000	0	0,000	5	0,070	0	0,000	1	0,200	0	0,000
V. PRES	7	0,032	0	0,000	1	0,040	1	0,040	5	0,070	0	0,000	0	0,000	0	0,000
DEJ	46	0,206	0	0,000	8	0,320	0	0,000	16	0,225	5	0,600	1	0,200	0	0,000
DAD	24	0,108	0	0,000	1	0,040	2	0,286	12	0,167	0	0,000	1	0,200	0	0,000
DPC	26	0,117	1	0,334	4	0,160	1	0,143	3	0,042	0	0,000	0	0,000	0	0,000
CPD	16	0,072	0	0,000	0	0,000	0	0,000	4	0,058	1	0,200	0	0,000	0	0,000
GEP	59	0,264	2	0,667	8	0,320	1	0,143	12	0,167	0	0,000	1	0,200	1	1,000
GCT	38	0,170	0	0,000	3	0,120	2	0,286	14	0,195	0	0,000	0	0,000	0	0,000
GF	1	0,005	0	0,000	0	0,000	0	0,000	1	0,014	1	0,200	1	0,200	0	0,000
GE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
N. BID	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
G. GAS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PNTE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TOTAL	224	1,000	3	1,000	25	1,000	7	1,000	72	1,000	5	1,000	5	1,000	1	1,000
X^2	—			3,801		7,388		8,306		28,282		12,538		30,957		2,809

Sigla	Código do projeto		Data do documento		Data do protocolo		Estados do Brasil		Executor do projeto		Fontes de recursos		Instituição de destino		Instituição de origem		Mutuário		Nome da consultora	
	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr
PRES	00	0,000	2	0,049	0	0,000	0	0,000	1	0,256	1	1,000	1	0,048	3	0,056	3	0,043	3	0,112
V. PRES	4	0,034	1	0,085	1	0,112	0	0,000	0	0,000	0	0,000	1	0,048	4	0,076	1	0,015	2	0,075
DEJ	26	0,219	4	0,098	0	0,000	1	0,500	1	0,256	0	0,000	5	0,239	10	0,186	21	0,300	4	0,149
DAD	10	0,085	13	0,318	0	0,000	0	0,000	0	0,000	0	0,000	6	0,286	9	0,167	6	0,086	3	0,112
DPC	14	0,118	2	0,049	1	0,112	0	0,000	0	0,000	0	0,000	0	0,000	4	0,075	6	0,086	1	0,038
CPD	8	0,068	3	0,074	3	0,334	0	0,000	0	0,000	0	0,000	1	0,048	7	0,130	2	0,029	1	0,038
GEP	29	0,244	6	0,147	0	0,000	1	0,500	0	0,000	0	0,000	3	0,143	7	0,130	21	0,300	10	0,371
GCT	28	0,236	10	0,244	4	0,445	0	0,000	2	0,500	0	0,000	4	0,191	9	0,167	10	0,143	3	0,112
GF	6	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,019	0	0,000	0	0,000
GE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
N. BID	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
G. GAS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PNTE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TOTAL	119	1,000	41	1,000	9	1,000	2	1,000	4	1,000	1	1,000	21	1,000	54	1,000	70	1,000	27	1,000
X^2		17,483		30,498		20,829		2,564		11,314		31,139		11,729		20,450		11,064		12,813

Sigla	Nome do projeto		N.º do documento		N.º do protocolo		N.º registro da consultora		Pessoa destinatária		Referências		Setor FINEP		Signatário		Tipo de documento		Valor		Variação nome projeto		Veículo da inf.		Σ f
	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	f	fr	
PRES	0	0,000	2	0,055	0	0,000	0	0,000	1	0,072	0	0,000	1	0,334	1	0,143	7	0,048	2	0,223	0	0,000	1	0,091	36
V. PRES	0	0,000	2	0,055	1	0,072	0	0,000	1	0,072	0	0,000	0	0,000	0	0,000	7	0,048	0	0,000	0	0,000	1	0,091	33
DEJ	4	0,400	3	0,074	1	0,072	0	0,000	3	0,216	1	0,125	1	0,334	0	0,000	30	0,203	0	0,000	4	0,250	1	0,091	148
DAD	1	0,100	6	0,158	0	0,000	0	0,000	3	0,216	0	0,000	1	0,334	3	0,429	21	0,142	3	0,334	0	0,000	3	0,273	104
DPC	1	0,100	6	0,158	1	0,072	0	0,000	0	0,000	0	0,250	0	0,000	0	0,000	15	0,129	1	0,112	5	0,188	1	0,091	71
CPD	1	0,100	1	0,027	4	0,286	0	0,000	1	0,072	1	0,125	0	0,000	2	0,286	10	0,068	0	0,000	0	0,000	0	0,000	50
GEP	2	0,200	8	0,211	0	0,000	11	1,000	3	0,215	2	0,250	0	0,000	0	0,000	28	0,190	0	0,000	8	0,500	3	0,273	167
GCT	1	0,100	10	0,264	7	0,500	0	0,000	2	0,143	2	0,250	0	0,000	1	0,143	25	0,169	5	0,534	1	0,063	1	0,091	142
GF	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	0	0,000	1	0,007	0	0,000	0	0,000	0	0,000	5
GE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
N. BID	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
G. GAS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PNTE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TOTAL	10	1,000	38	1,000	14	1,000	11	1,000	14	1,000	8	1,000	3	1,000	7	1,000	148	1,000	9	1,000	16	1,000	11	1,000	755
X^2		3,239		11,043		27,606		33,551		8,149		3,642		12,569		19,007		22,414		22,504		9,984		7,701	–

Admitindo-se que os Departamentos exercem atividades diversificadas, a metodologia mais indicada para identificação do núcleo de campos, comum a todos os Departamentos, seria o teste de significância do x^2, medida estatística capaz de estabelecer os limites de discrepância entre diferentes classes, a partir da comparação entre a freqüência probabilisticamente esperada e aquela obtida. Essa medida pode ser expressa através da seguinte fórmula:

$$x^2 = \frac{(fe - ft)^2}{ft}$$

sendo fe a freqüência empírica ou obtida e ft a freqüência teórica ou esperada.
Para efeitos da presente metodologia, para 12 G.L. foi adotado o nível de significância 0,01, rejeitando-se por conseguinte, os valores acima de 26,2 conforme distribuição do x^2_P (9), uma vez que se buscavam os campos de uso generalizado e não aqueles estreitamente relacionados a Departamentos específicos.

Os resultados (Tabelas 4-5) apresentaram 23 campos aceitos e 8 campos rejeitados por não serem de importância no consenso geral e sim de interesse particular de algum Departamento. Observando a tabela 3, verifica-se que 2 daqueles campos aceitos apresentaram-se homogêneos, porém em torno de zero, justificando sua exclusão. São eles: classificação do projeto e Estados do Brasil.

Os campos aceitos são: nome do projeto, referência, agência do AF, pessoa destinatária, agente financeiro, veículo da informação, anexos, variações do nome do projeto, número do documento, mutuário, executor do projeto, signatário, instituição de destino, data do protocolo, valor, tipo de documento e assunto.
Tais campos representam o núcleo comum a todos os Departamentos e constituem o conjunto mínimo capaz de satisfazer as necessidades básicas dos usuários. Os restantes seriam incorporados ao banco de dados conforme as possibilidades da Empresa, na medida em que a satisfação das necessidades específicas de cada Departamento se constituir uma meta.

TABELA 4: CAMPOS ACEITOS SEGUNDO O TESTE DE SIGNIFICÂNCIA * (RIO DE JANEIRO, JULHO DE 1974)

Nome dos Campos	x^2
Estados do Brasil	2,354
Classificação do projeto	2,809
Nome do projeto	3,239
Referências	3,642
Agência do AF	5,001
Pessoa destinatária	5,149
Agente financeiro	7,288
Veículo da informação	7,701
Anexos	8,306
Variações do nome do projeto	9,984
Número do documento	11,043
Mutuário	11,084
Executor do projeto	11,314
Instituição de destino	11,729
Setor FINEP	12,569
Nome da consultora	12,813
Código do projeto	17,483
Signatário	19,007
Instituição de origem	20,456
Data do protocolo	20,829
Valor	22,304
Tipo de documento	22,414
Assunto	25,262

* 12 G.L. e nível de significância 0,01

Fonte: Pesquisa da autora no Arquivo da FINEP

TABELA 5: CAMPOS REJEITADOS SEGUNDO O TESTE DE SIGNIFICÂNCIA * (RIO DE JANEIRO, JULHO DE 1974)

Nome dos Campos	x^2
Número do protocolo	27,606
Data do documento	30,478
Fonte de recursos	31,139
Número de registro da consultora	32351
Cargo do signatário	50,937
Cargo do destinatário	52,538

* 12 G.L. e nível de significância 0,001
Fonte: Pesquisa da autora no Arquivo da FINEP

RESULTADOS

Com relação à FINEP, se se considerar que há atividades comuns como há também as específicas de cada setor, parece que a combinação das duas metodologias — teste do x^2 corrigido pela ordenação simples de freqüência — é o recomendado, para maior segurança.
Assim sendo, seriam considerados para inclusão no banco de dados os campos aceitos pelo x^2 e aqueles rejeitados que estivessem situados dentre os 16 campos (Tabela 1) de maior freqüência na série (fr = 0,752). Desses 24 componentes aceitos, 5 poderiam ser ainda descartados peks razões que se seguem:

— Setor FINEP: porque a Empresa é ainda relativamente pequena e todos os funcionários conhecem os diferentes programas da instituição, sendo capazes de com eles relacionar os projetos;

— nome do projeto: existe uma parte do banco de dados destinada ao cadastro, no qual todas as características dos projetos são registradas. Portanto, bastaria relacionar o arquivo DOCUMENTOS com o CADASTRO, para que as informações gerais se tornassem disponíveis;

— variações do nome do projeto: ficaria melhor no CADASTRO, onde este campo deveria ser introduzido;

— mutuário: Já existe no CADASTRO;

— nome da consultora: já existe no CADASTRO, além de poder ser substituído economicamente pelo número de registro da consultora.

Os 19 campos restantes parecem satisfazer às necessidades básicas de todos os Departamentos: tipo de documento, código do projeto, assunto, instituição de origem, data do documento, número do documento, agente financeiro, instituição de destino, número do protocolo, veículo da informação, número de registro da consultora, pessoa destinatária, signatário, data do protocolo, valor, referências, anexos, executor do projeto e agência do AF.
A adoção do sistema integral, visando a atingir o desempenho ótimo, ou a adoção de apenas parte dos campos ficará a critério da Administração da Empresa, segundo sua política interna.

CONCLUSÕES GERAIS

A partir da presente pesquisa conclui-se que:

— a escolha dos componentes do banco de dados deve ser determinada em função de seu peso (fr), uma vez que esse é indicativo da preferência do usuário;

— só é economicamente justificável a definição dos campos a partir de sua ocorrência na documentação, quando esta freqüência coincide com a de uso, pois a alta freqüência nos documentos está diretamente relacionada com o alto custo da inclusão dos campos no sistema;

— o número de campos a serem eleitos após a duração dos mais importantes dependerá da política da Empresa quanto à canalização de recursos para o sistema;

— duas metodologias, baseadas nas necessidades expressas dos usuários, podem ser utilizadas para identificação dos campos fundamentais para o sistema atingir seu desempenho ótimo na recuperação de informações, dependendo das características dos usuários:
  — ordenação das freqüências relativas dos campos e eleição, a partir dos mais freqüentes, quando o grupo de usuários é homogêneo;
  — teste do x^2, quando as atividades e necessidades dos grupos diferem-se entre si.

Em uma situação específica, qualquer das duas metodologias sendo adotada, seria recomendável que periodicamente fossem reavaliados os interesses dos usuários do sistema, a fim de corrigir o modelo proposto a partir da pesquisa inicial.

CITAÇÕES BIBLIOGRÁFICAS

(1) DAMMERS, H. F. Information management systems: some views on problems and potentialities. In: DATA ORGANIZATION FOR MAINTENANCE AND ACCESS CONFERENCE, Keele, April 1970. *Papers.* Keele, The University, 1970.

(2) ESPOSEL, José Pedra Pinto. Editorial. *Arquivos & Administração,* 2 (2):5, ago. 1974.

(3) VIEIRA, A. S. *Metodologia para definição de campos em bancos de dados.* Rio de Janeiro, 1974. 52 p.

(4) SOUSA, Flávio Pereira. *Introdução* à *recuperação da informação.* /A ser publicada ainda em 1974 pelo convênio MEC/PUC/.

(5) SALTON, Gerard. *Automatic information organization and retrieval.* New York, Mac-Graw-Hill, 1968. 514 p.

(6) CLEVERDON, Cyril. Information and its retrieval. *Aslib Proceedings,* 22 (11): 546, Nov. 1970.

(7) LANCASTER, F. Wilfrid. MEDLARS: report on the evaluation of its operating efficiency. *American Documentation,* 20(2): 119-42, Apr. 1969.

(8) HALD. *Statistical tables and formulas.* New York, Willey, 1952. p. 96.

(9) SPIEGEL, M. R. *Estatística.* Tradução de Pedro Consentino, São Paulo, McGraw-Hill do Brasil, 1974. 580 p.

BIBLIOGRAFIA CONSULTADA

CUNHA, S, E. *Estatística Descritiva (na Psicologia e Educação)* Rio de Janeiro, Forense /s. d. /243 p.

ENGELS, R. W. A tutorial on data-base organization; TR 00.2004. In: IBM. *Data base concepts; education guide.* New York, 1972.

FARRADANE, J. The evaluation of information retrieval systems. *Journal of Docurruintation,* 30 (2): 195-209, June 1974.

GELLER, S. B. Archival data storage. *Datamation,* 20 (10): 72-80, Oct. 1974.

KEMP. D. A. Relevance, partinence and information system development. *Information Storage and Retrieval,* 10(2):37-47, Feb. 1974.

KING, D. W. & BRYANT, E. C. *The evaluation of information services and products.* Washington, Information Resources Press, 1971. 306 p.

KONIGOVÁ, M. Mathematical and statistical methods of noise evaluation in a retrieval system. *Information Storage and Retrieval,* (6): 437-44, May 1971.

LANCASTER, F. W. *Information retrieval systems; characteristics, testing, and evaluation.* New York, J. Wiley, 1968. 222 p.

———. & FAYEN, E.G. *Information retrieval on-line.* Los Angeles, Melville Publishing, 1973. 597 p.

LANDAU, H. The proliferation of machine-readable data bases: current problems. *Drexel Library Quarterly,* 8(1): 63-9, Jan. 1972.

MARTYN, J. & VICKERY, B.C. The complexity of modelling of information systems. *Journal of Documentation,* 26(3): 204-20, Sept. 1970.

NICK, E. & KELLNER, S.R.O. *Fundamentos de estatística para as ciências do comportamento.* Rio de Janeiro, Renes, 1971. 312 p.

RIEGER, M. Le role des archives dans l'administration. *Bulletin de l'Unesco pour les Bibliothèques,* 27(l):43-5, Jan./Fev. 1973.

SAFFADY, W. A university archives and records management program: some operational guidelines. *College & Research Libraries, 35* (3):204-10, May 1974.

SALTON, G. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval,* 6(l):29-44, May 1970.

& YANG, C. S, On the specification of term values in automatic indexing. *Journal of Documentation,* 29(4): 351-72, Dec. 1973.

SEELY, B. J. Indexing depth and retrieval effectiveness. *Drexel Library Quarterly,* 8(2):201-8, Apr. 1972.

TAKAHAMA, T. A model for a document retrieval system. *Information Storage and Retrieval,* 9(3):143-63, Mar. 1973.

VICKERY, B.C. *Information systems.* London, Butterworths, 1973. 350 p.

WILSON, P. Situational relevance. *Information Storage and Retrieval,* 9(8):457-71, Aug. 1973.

ABSTRACT

Based on data collected at FINEPs (Financiadora de Estudos e Projetos) Archives and having the aim of building up a data base on typical documents related to project administration, two alternative methodologies were designed, using statistical measures, to define which fields of information should be used at the system. The first methodology is based on the frequency order of the fields, according to their frequency at users *question, and should be useful when the users have common interests and activities. The second methodology — the* x^1 *test — would be suitable when users have different interests and activities.*