

SISTEMA ESPECIALISTA PARA VALIDAÇÃO DE BASES ARUANDA

Angela Maria Bosígio

Sedes/Arinf/Disul

292-4100 ramal 592

SERVIÇO FEDERAL DE PROCESSAMENTO DE DADOS - SERPRO

Rua da Lapa, nº 40 - CEP: 20.000

Rio de Janeiro - BRASIL

Marcia Pacitti Teles

Sfgts/Digs/Deat-2

292-4100 ramal 208

RESUMO

O Sistema Especialista destina-se a verificar se uma determinada base definida para o Sistema Aruanda, está obedecendo às normas de adequação da arquitetura Aruanda. Além disto, são geradas recomendações para que o projetista possa melhorar e/ou corrigir a definição da base fornecida para o sistema.

ABSTRACT

The purpose of the expert system is to verify if a given base defined to the Aruanda System is suitable to the norms and adequate to the Aruanda architecture. Besides that, the expert system generates recommendations to the projetist so that he can optimize and/or correct his original definition of the base furnished to the system.

1. ORIGEM HISTÓRICA DO PROJETO

O Aruanda é um sistema de gerenciamento de Banco de Dados capaz de armazenar, administrar e distribuir um conjunto teoricamente não limitado de bases de dados com formatos e conteúdos diversificados, consultados e acessados segundo algum critério de seleção.

A falta de vivência da cultura Aruanda acarreta dificuldades no projeto de incorporação de uma nova base. Verificamos que seria necessário investir na disseminação desta cultura. A nossa experiência revela que o tempo de maturação de um projetista de bases não sofisticadas é bastante longo.

O que nos pareceu uma solução interessante foi criar um Sistema Especialista para ajudar ao projetista validar sua base. O sistema contém na sua base de conhecimento as regras de validação necessárias para a incorporação de uma nova base ao Aruanda. Na base de fatos do sistema, a estrutura das bases são armazenadas de modo que possam ser recuperadas e consultadas pelo projetista.

O desenvolvimento deste projeto teve início em Março de 1987. No momento ele se encontra em fase final de testes.

2. DEFINIÇÃO ARUANDA

2.1 Características Gerais

O Aruanda é um sistema de acesso a bases de dados, que tem como objetivo principal ser um sistema de disseminação de informação de uso geral e público.

Foi desenvolvido no Serpro em 1978, utilizando a linguagem PL/1, e os sistemas Adabas e Intercomm.

Para se alcançar o objetivo principal do Aruanda sua filosofia de recuperação incorpora os seguintes quesitos básicos:

a) Generalidade

O Aruanda não se destina a lidar apenas com um determinado tipo de informação. É capaz de tratar arquivos de dados com os mais diversos conteúdos lógicos e formatos físicos, ou seja, é capaz de acessar diferentes bases de dados com uma única linguagem.

b) Conversacionalidade

A comunicação usuário / sistema remete o usuário de uma situação onde ele dispõe apenas de seu vocabulário habitual para fazer a consulta, até uma outra situação onde ele finalmente recupera um registro ou um campo contendo a informação desejada.

A filosofia de recuperação supõe para isto, tanto a utilização de um vocabulário "livre" (o próprio conteúdo da informação) como de um vocabulário controlado cujos termos indexam a informação e funcionam como chaves de recuperação.

A fim de atender a estes quesitos básicos, foi definida uma determinada estrutura lógica para o sistema, e também criada uma linguagem de acesso a ser usada para recuperação em todas as bases de dados componentes do Aruanda, linguagem esta chamada LA (Linguagem Aruanda).

2.2 Arquitetura Aruanda

Todo projeto e funcionamento do Aruanda se baseia, portanto, numa estrutura lógica de dados, constituída dos seguintes elementos:

BASE DE DADOS
ARQUIVO DE DADOS
UNIDADE DE INFORMAÇÃO
ÍNDICE DE SELEÇÃO
DICIONÁRIO
DESCRITOR
VARIÁVEIS RECUPERÁVEIS
FORMATO DE IMPRESSÃO

Onde,

- a) UNIDADE DE INFORMAÇÃO é o conjunto de dados que constitui a informação básica a ser recuperada.
- b) ARQUIVO DE DADOS é o conjunto onde cada elemento é uma unidade de informação.
- c) ÍNDICE DE SELEÇÃO - os índices de seleção são chaves de pesquisa que são utilizadas na recuperação do Aruanda. Na recuperação da informação existem dois métodos de recuperação. O primeiro utiliza um critério de seleção acessando o arquivo de dados diretamente. O segundo utiliza um critério de seleção indireto, acessando um outro conjunto de informação (Dicionário), que de alguma forma estará indexando as informações do arquivo de dados. Por este motivo, os índices de seleção são classificados em dois tipos segundo o método de recuperação no qual ele será utilizado.
- d) DICIONÁRIO é um conjunto determinado de todos os descritores que compõem um vocabulário controlado. Quando o dicionário é utilizado o meio de acesso as informações é indireto.
- e) DESCRITOR é cada elemento do dicionário. Cada descritor é identificado padronizadamente por seu código e por sua descrição. Cada descritor está associado a uma ou mais unidades de informação do arquivo de dados.
- f) VARIÁVEIS RECUPERÁVEIS é o conjunto de itens da unidade de informação escolhidos para serem utilizados no método de seleção direto.
- g) FORMATO DE IMPRESSÃO é um subconjunto de dados da Unidade de Informação escolhidos para exibição ao usuário.
- h) BASE DE DADOS ARUANDA é constituída pelos elementos definidos acima, ou seja, a tripla (Arquivo de Dados, Conjuntos de Índices de Seleção, Formatos de Impressão).

2.3 Validações Necessárias

2.3.1 Definição da Base Aruanda para fins de Validação

A fim de validar a arquitetura de uma base, o sistema especialista deve analisar as entidades contidas na sextupla (Arquivo de Dados, Conjunto Índices Dicionário, Conjunto Índices Variáveis Recuperáveis, Formatos de Impressão, Dicionário, Associação dos Índices), que foi definida em função da arquitetura vista no item 2.2.

Os índices foram divididos da seguinte forma :

- Entidade Conjunto Índices Dicionário, constituída de chaves de recuperação, que estão indexando indiretamente itens da Unidade de Informação, onde para cada índice existe um Dicionário associado.
- Entidade Conjunto Índices Variáveis Recuperáveis, constituída de chaves de recuperação, que estão associadas diretamente

a itens da Unidade de Informação, escolhidos como índices.

Esta divisão foi efetuada pois o Aruanda trata de modo diferente estes dois tipos de índices, o que acarreta validações distintas para eles.

Além disto, ao Conjunto Índices Dicionário está associada a entidade Dicionário.

Criamos uma outra entidade que denominamos de ASSOCIAÇÃO DOS ÍNDICES, que indica a associação entre um elemento do Conjunto de Índices Variáveis Recuperáveis e a correspondente variável na Unidade de Informação, ou a associação que representa a indexação entre um elemento do Conjunto de Índices Dicionário e um item da Unidade de Informação.

A validação a qual nos propuzemos, foi verificar se as entidades de uma dada sextupla satisfazem aos requisitos necessários ao funcionamento do Sistema Aruanda.

2.3.2 Regras de Validação

O conjunto de regras a seguir, usadas na validação de uma Base de Dados Aruanda, representa a base de conhecimento do Sistema. O número de regras não depende de cada Base, pois são regras para verificar se uma Base de Dados qualquer satisfaz a arquitetura Aruanda.

Definiremos inicialmente a linguagem L, que representa as regras de validação Aruanda, como uma linguagem de primeira ordem, com notação

$\langle V, PL, FL, CL \rangle$ onde V, PL, FL, CL são conjuntos assim constituídos:

V = conjunto de variáveis

F = { campo, listacampo, listaíndices, listavarecus,

L listaformatos, estruturaformato, associação, listassociação, listanomes... }

conjunto de funções de L

C = N U { cadeia de caracteres }

L U { alfa, num, descritor, periódico }

conjunto de constantes de L

P = { $\in$, inteiro, =, $\leq$, $\geq$, ehargumento } U

L { formato, atributo1, atributo2, atributo3, base_aruanda, unidade_de_informação, arquivo_de_dados, índices_em_lu, ... }

conjunto de predicados de L

Apresentaremos a seguir os termos de L, que representam os elementos da sextupla Aruanda.

1) Arquivo de Dados, notação LU

LU, termo de L é assim definido:

```
listacampos(campo(n , t , f , at1 , at2 , at3 ),
              1 1 1      1      1      1
              campo(n , t , f , at1 , at2 , at3 ),    ...,
              2 2 2      2      2      2
              campo(n , t , f , at1 , at2 , at3 ))
              m m m      m      m      m
```

onde $n \in V$, $t \in V$, ... etc

$1 \leq i \leq m$ $1 \leq i \leq n$

Observe que o termo $\text{campo}(n , t , f , at1 , at2 , at3)$ está

associado ao elemento unidade de informação do Aruanda.

2) Conjunto Índices Dicionário, notação LID

LID, termo de L, é assim definido:

$listaíndices(n_1, n_2, \dots, n_m)$, onde $n_i \in V$
 $1 \leq i \leq m$

3) Conjunto Índices Variáveis Recuperáveis, notação LVR

LVR, termo de L, é assim definido:

$listavarecus(n_1, \dots, n_m)$, onde $n_i \in V$
 $1 \leq i \leq m$

4) Formatos de Impressão, notação LFR

LFR, termo de L, é assim definido:

$listaformatos(estruturaformato(m_1, listanomes(n_1, \dots, n_{1p})),$
 $estruturaformato(m_2, listanomes(n_2, \dots, n_{2p})),$
 $\dots$
 $estruturaformato(m_l, listanomes(n_l, \dots, n_{lp})))$
 $1 \qquad \qquad \qquad l \qquad \qquad \qquad lp$

onde $m_j \in V$, $n_k \in V$, com $1 \leq j \leq l$, $1 \leq k \leq p$
 $1 \leq i \leq l \qquad jk$

5) Dicionário, notação DIC

DIC, termo de L, é assim definido:

$regdic(campocodigo_1(n_1, t_1, f_1, at1_1, at2_1, at3_1),$
 $campodescrição1(n_2, t_2, f_2, at1_2, at2_2, at3_2),$
 $campodescrição2(n_3, t_3, f_3, at1_3, at2_3, at3_3),$
 $campovps_4(n_4, t_4, f_4, at1_4, at2_4, at3_4),$
 $campotg_5(n_5, t_5, f_5, at1_5, at2_5, at3_5),$
 $campotc_6(n_6, t_6, f_6, at1_6, at2_6, at3_6),$
 $campotr_7(n_7, t_7, f_7, at1_7, at2_7, at3_7),$
 $camposn1_8(n_8, t_8, f_8, at1_8, at2_8, at3_8),$
 $\dots$
 $camposn6_m(n_m, t_m, f_m, at1_m, at2_m, at3_m))$
 $m \quad m \quad m \quad m \quad m \quad m$

6) Dicionários, notação LDIC

LDIC, termo de L, é assim definido:

$listaregdic(DIC_1, DIC_2, \dots, DIC_n)$
 $1 \qquad 2 \qquad \qquad n$

onde DIC, é termo de L definido no item 5.
 $1 \leq i \leq n$

7) Associação Índice, notação AID

AID, termo de L, é assim definido:

$associação(m, campo(n, t, f, at1, at2, at3))$

8) Associação dos índices, notação LAID

LAID, termo de L, é assim definido:

$listaassociação(AID_1, AID_2, \dots, AID_n)$

onde AID é termo de L definido no item 7.

$1 \leq i \leq n$

Uma base de dados Aruanda, será definida pelo seguinte predicado:

$base_aruanda(LU, LID, LVR, LFR, LDIC, LAID)$, onde LU, LID, LVR, LFR, LDIC, LAID são termos de L.

Diz-se, que uma base de dados Aruanda está "correta", se dado LU, LID, LVR, LFR, LDIC, LAID, conseguirmos provar através das regras de validação, que o predicado $base_aruanda(LU, LID, LVR, LFR, LDIC, LAID)$ é verdadeiro.

Mostraremos, a seguir, um pequeno subconjunto de critérios utilizados na validação de uma base Aruanda.

Regras de Validação

I) Dado LU, LID, LVR, LFR, LDIC, LAID

$\wedge$ arquivo_de_dados(LU)

$\wedge$ indices_em_lu(LID, LVR, LDIC, LAID)

$\wedge$ conj_formato_impressão(LFR)

$\Rightarrow base_aruanda(LU, LID, LVR, LFR, LDIC, LAID)$

II) Dado $LU = lista_campo(campo(n_1, t_1, f_1, at1_1, at2_1, at3_1), \dots, campo(n_i, t_i, f_i, at1_i, at2_i, at3_i))$

$\forall n_i, t_i, f_i, at1_i, at2_i, at3_i, 1 \leq i \leq i$

$\wedge unidade_informação(campo(n_i, t_i, f_i, at1_i, at2_i, at3_i))$

$\Rightarrow arquivo_de_dados(LU)$

III) Dado $campo(n, t, f, at1, at2, at3)$

$\wedge (n, (cadeia \text{ de caracteres}))$

$\wedge inteiro(t)$

$\wedge formato(t)$

$\wedge atributo1(at1)$

$\wedge atributo2(at2)$

$\wedge atributo3(at3)$

$[(at2, multiplo) \Rightarrow inteiro(at3)]$

$\Rightarrow unidade_de_informação(campo(n, t, f, at1, at2, at3))$

IV) $\forall f [(f, alfa) \vee (f, num) \Rightarrow formato(t)]$
 $\forall at1 [(at1, descriptor) \vee (at1, \emptyset) \Rightarrow atributo1(at1)]$
 $\forall at2 [(at2, multiplo) \vee (at2, \emptyset) \Rightarrow atributo2(at2)]$
 $\forall at3 [(at3, inteiro) \vee (at3, \emptyset) \Rightarrow atributo3(at3)]$

V) Dado LID,LVR,LDIC,LAID

conjunto_indice_dicionário(LID,LDIC,LAID)
 $\vee$ conjunto_indice_varecu(LVR,LAID)

=> indices_em_lu(LID,LVR,LDIC,LAID)

VI) Dado LU,

LID=listaindices($n_1, \dots, n_m$),

LDIC=listaregdic(DIC $_1, \dots, DIC_x$)

LAID=listaassociação(AID $_1, \dots, AID_k$)

$\forall n \in \exists DIC,$

$1 \leq i \leq m \quad 1 \leq p \leq x$

$\exists AID = associação(n, campo(m, t, f, at1, at2, at3))$

$1 \leq q \leq k \quad 1 \leq j \leq J \quad 1 \leq J \leq J$

[dicionário_para_indice_em_lu(DIC, n)

($\wedge$ associação_para_indice_dic_em_lu(AID $_p$))]]

=> conjunto_indice_dicionário(LID,LDIC,LAID)

Obs: Cada $\exists$ "termo de L", está sendo usado para simplificar. Na verdade, teríamos que usar vários $\exists$ s para cada variável, constituinte do termo quantificado.

VII) Dado LU,

LVR=listavarecus($v_1, v_2, \dots, v_k$),

LAID=listaassociação(AID $_1, \dots, AID_m$)

$\forall v \in \exists AID = associação(v, campo(n, t, at1, at2, at3))$

$1 \leq i \leq k \quad 1 \leq j \leq m \quad 1 \leq k \leq k \quad 1 \leq k \leq k$

[associação_para_indice_varecu_em_lu(AID))]]

=> conjunto_indice_varecu(LVR,LAID)

3. IMPLEMENTAÇÃO DO SISTEMA

3.1 Enfoque Adotado na Implementação das Entidades Principais

A linguagem adotada para desenvolver o sistema foi o Prolog. O projetista deve fornecer ao sistema as informações relevantes a cada entidade. O sistema irá requerer do projetista a configuração de cada entidade, pedindo os campos de entrada de cada entidade separadamente. Com isso, o projetista passa a compreender melhor as entidades básicas do Aruanda.

A primeira entidade a que chamamos de Unidade de Informação contem os campos básicos que podem ser recuperados no Sistema Aruanda. Para cada campo o projetista deve fornecer seu nome, tamanho, formato e atributos. Os atributos servem para designar

se um campo é descritor, isto é, se ele pode ser utilizado como chave de pesquisa, ou ainda se é um campo múltiplo, ou seja, ocorre várias vezes dentro de uma unidade de informação. Uma lista de estruturas chamada LU, é construída a medida que os campos são fornecidos. Cada campo possui a seguinte definição :

campo (X,T,F,At1,At2,At3)

onde:

X - nome do campo

T - tamanho

F - formato

At1 - informa se é campo chave (será referenciado como Atributo 1)

At2 - informa se é um campo múltiplo (será referenciado como Atributo 2)

At3 - informa o número de ocorrências do campo múltiplo (será referenciado como Atributo 3)

A segunda entidade representa os índices que são utilizados para recuperação de Informação no Sistema Aruanda. Estes índices são agrupados em dois conjuntos distintos, como visto anteriormente.

Os elementos do primeiro grupo são os índices chamados de variáveis recuperáveis.

Uma lista chamada LVR é construída, contendo os nomes das variáveis recuperáveis, a medida que estes índices são fornecidos.

Os elementos do segundo grupo são chamados de índices dicionários. Estes índices são representados no sistema como uma lista de estruturas, denominada LID, e construída a medida que o projetista fornece estes índices. Cada elemento desta lista foi assim representado :

estrudic (X,Ndic)

onde:

X - nome do índice dicionário

Ndic - nome do arquivo dicionário associado

Para o Aruanda, como já foi visto, a um índice dicionário está associado a entidade Dicionário. Esta terceira entidade é associada explicitamente quando definimos o índice dicionário.

A entidade dicionário esta definida no sistema da seguinte forma:

Dic (M,Ndic,descriptor_aruanda(L))

onde:

M - nome da base

Ndic - nome do dicionário

descriptor_aruanda - estrutura que compreenderá uma lista de itens, necessários para que este índice seja tratado no Aruanda.

A estrutura descriptor_aruanda poderá compreender alguns ou todos dos seguintes itens:

- .campo_cod - campo que identifica um descritor Aruanda
- .campo_desc1 - campo que será necessário para descrição do conteúdo da ocorrência de um índice (do descritor Aruanda)
- .campo_desc2 - continuação da descrição acima
- .campo_vocab - necessário para a descrição dos vocábulos pertinentes de um descritor Aruanda

- .campo_tg - mostra se existe alguma relação hierárquica entre os descritores Aruanda em um índice
- .campo_tc - mostra se existe alguma relação de associação entre os descritores Aruanda em um índice
- .campo_tr - mostra se existe alguma relação de equivalência entre os descritores Aruanda em um índice

Cada uma das estruturas acima contem os seguintes elementos: X-nome, T-tamanho, F-formato, At1-atributo1, At2-atributo2, At3-atributo3 igual aos argumentos que definem um campo da Unidade de Informação.

A quarta entidade representa os formatos de exibição definidos para uma Base Aruanda. Esta entidade é representada através de uma lista de estruturas, denominada "estrufor", onde cada elemento é definido por:

estrufor (Nformato , Lista_nomes)

onde:

Nformato - identifica o nome do formato

Lista_nomes - uma lista que contem os nomes dos campos pertencentes ao formato

Existem dois tipos de formatos. O formato estático é definido quando todos os campos de Lista_nomes são nomes de campos da dita LU. Caso algum campo não pertença à LU, será denominado formato dinâmico.

3.2 Validação

A validação é realizada em cada entidade separadamente. Este enfoque é importante, pois quando existir uma falha detectada, e não for possível continuar validando aquela entidade, passamos a validar a próxima entidade. Veja bem, isto só é realizado se for possível continuar a validação das outras entidades, ou dela própria.

Verificamos que em certos pontos da validação, com o nosso conhecimento do Aruanda, seria oportuno gerar recomendações que possibilitariam ao projetista melhorar o desempenho de recuperação das informações na Base. Em termos da estrutura Aruanda, o que foi fornecido poderá estar correto, porém uma recomendação será emitida para que o usuário possa avaliar se uma outra estrutura não seria mais eficiente. Outras recomendações poderão implicar num alerta para problemas que poderão surgir na recuperação da informação, não implicando que a estrutura seja modificada.

3.3 Inserção de Fatos

A partir do momento que uma determinada base é validada pelo sistema e nenhum diagnóstico de erro é enviado, são inseridos fatos na base de fatos do Prolog, de modo que apenas estes fatos permaneçam após a validação. Os fatos armazenados são aqueles que representam as entidades já definidas anteriormente, vinculadas à base validada.

4. UTILIZAÇÃO DO SISTEMA

A entrada dos dados necessários para que a validação seja iniciada é executada antes do início da validação. A única entidade que é requisitada durante a validação é a que representa a entidade dicionário, caso o projetista necessite que seu índice seja dicionarizado.

Os campos são requisitados um de cada vez e as informações são indagadas explicitamente ao projetista de forma que este não omita nenhum dado relevante.

As mensagens decorrentes de falhas na validação ou de recomendação são exibidas na tela no momento em que estas são detetadas, porém o processo de validação continua até o final.

Como já foi visto anteriormente, a partir do momento que uma base é validada sem nenhum diagnóstico de falha, são inseridos fatos na base de fatos que representam as entidades associadas à base Aruanda em questão.

Estes fatos são depositados num arquivo para que, a qualquer momento, um projetista do Aruanda possa consultar uma imagem de seu próprio banco, que compreende todas as entidades de todas as bases já implantadas.

Isso proporciona uma ferramenta de apoio muito interessante para os projetistas. Por exemplo, caso uma nova base necessite de um índice dicionário que já tenha sido utilizado anteriormente para outras bases, ela pode ser consultada da seguinte forma:

?-Arq_dic (Base , nome_indice , Dic).

Com isso, todas os Arq_dics para onde este índice já tenha sido definido serão instanciados, e o projetista conhecerá todas as descrições dos índices já existentes.

5. CONCLUSÃO

Esta foi a nossa primeira experiência no desenvolvimento de um sistema especialista, e achamos que foi muito gratificante. O sistema irá ajudar a preservar e disseminar a cultura Aruanda, que foi o nosso principal objetivo, facilitando ao projetista a utilização do Sistema Aruanda.