

OpenRecLink



Guia do Usuário

Rio de Janeiro
2012

Sumário

Créditos	1
Introdução.....	2
Arquivo	3
Ferramentas	3
Arquivo.....	3
Visualiza.....	3
Estrutura.....	3
Reinicia.....	3
Índice.....	4
Movimenta.....	4
Anexa.....	4
Exporta.....	4
Importa.....	5
Importa dbf.....	5
Sai.....	5
Log Atividade	6
Sobre	6
Sai	6
Copia.....	7
Configura.....	7
Executa.....	10
Padroniza	11
Configura	11
Executa	15
Relaciona	16
Configura	16
Gera freq	20
Analisa freq.....	21
Gera matrizes	22
Calcula	23
Executa	24
Reduz pares	25
Marca/Desmarca	26
Exporta SQL	27
Deduplica.....	28
Configura	28
Executa	30
Examina	31
Combina	32
Examina	32
Executa	33
Associa	34
Configura.....	34
Executa.....	38
Notas técnicas.....	39
ANEXO: GPL Versão 3.....	41
GNU GENERAL PUBLIC LICENSE.....	41
Preamble.....	41
TERMS AND CONDITIONS.....	42
0. Definitions.....	42
1. Source Code.....	43

2. Basic Permissions.....	43
3. Protecting Users' Legal Rights From Anti-Circumvention Law.....	44
4. Conveying Verbatim Copies.....	44
5. Conveying Modified Source Versions.....	44
6. Conveying Non-Source Forms.....	45
7. Additional Terms.....	47
8. Termination.....	48
9. Acceptance Not Required for Having Copies.....	48
10. Automatic Licensing of Downstream Recipients.....	48
11. Patents.....	49
12. No Surrender of Others' Freedom.....	50
13. Use with the GNU Affero General Public License.....	50
14. Revised Versions of this License.....	50
15. Disclaimer of Warranty.....	51
16. Limitation of Liability.....	51
17. Interpretation of Sections 15 and 16.....	51
END OF TERMS AND CONDITIONS.....	51

Créditos

Concepção e programação:

Kenneth Rochel de Camargo Jr.
Cláudia Medina Coeli

Implementação de rotinas de estimação de parâmetros:

Washington Leite Junger

Apoio ao desenvolvimento do programa:

UERJ
UFRJ
CNPq
FAPERJ

Introdução

O relacionamento de bancos de dados vem sendo crescentemente utilizado na pesquisa em saúde. O método determinístico pode ser empregado quando um campo identificador unívoco (ex. CPF, número de cartão de saúde) encontra-se presente nos bancos a serem relacionados. Na ausência deste identificador, o relacionamento pode ser executado empregando-se o método probabilístico (*probabilistic record linkage*). Este último baseia-se na utilização conjunta de campos comuns presentes em ambos os bancos de dados (ex. nome, data de nascimento), com o objetivo de identificar o quanto é provável que um par de registros se refira a um mesmo indivíduo.

No relacionamento probabilístico são executados um conjunto de processos, a saber: (1) a utilização de rotinas para a padronização dos campos comuns a serem empregados no relacionamento (ex. quebra do campo nome em seus componentes e a formatação de campos data); (2) a blocagem (*blocking*), que consiste na criação de blocos lógicos de registros dentro dos arquivos a serem relacionados, permitindo que a comparação entre registros se faça de uma forma mais otimizada; (3) a aplicação de algoritmos para a comparação aproximada de cadeias de caracteres, que levam em consideração possíveis erros fonéticos e de digitação (ex. Manoel e Manuel seriam reconhecidos como iguais); (4) o cálculo de escores, que sumarizam o grau de concordância global entre registros de um mesmo par; (5) a definição de limiares para a classificação dos pares de registros relacionados em pares verdadeiros, não pares e pares duvidosos; (6) a revisão manual dos pares duvidosos visando a classificação dos mesmos como pares verdadeiros ou não pares; (7) a verificação de duplicidade de registros em arquivos.

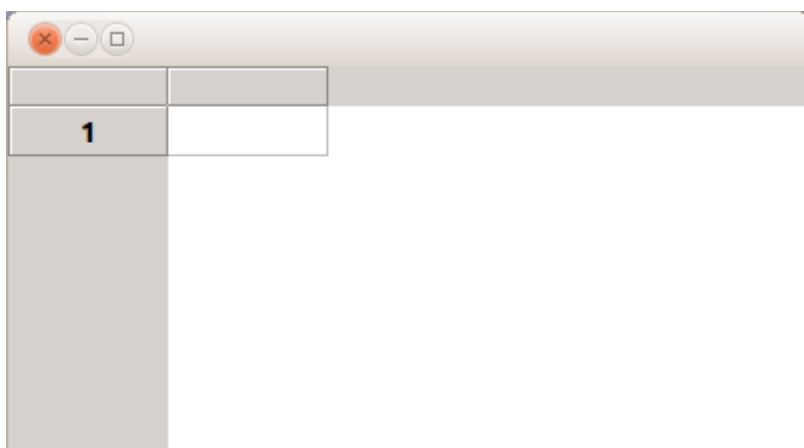
O programama **OpenRecLink** implementa várias rotinas de processamento de arquivos, em especial a associação com base na técnica de relacionamento probabilístico de registros. Este documento apresenta sua operação.

O programa e seu código fonte estão disponíveis sem custo para quaisquer interessados, nos termos da licença GPL versão 3 (ver ao final deste documento).

Arquivo

Ferramentas

Este módulo agrupa várias funções utilitárias de acesso aos arquivos de dados utilizados pelo programa (ver nota técnica ao final deste manual). Ao escolher esta opção, será aberta uma janela simples, contendo o menu e uma grade inicialmente vazia:



Arquivo

Este menu concentra as funções do módulo que operam em um único arquivo.

Visualiza

Esta é a função mais básica, que abre um arquivo e mostra seu conteúdo sob a forma de tabela na grade da janela.

A screenshot of a software window titled 'demolinkSCORE'. The window displays a table with five columns: an index column, 'COMPREC', 'REFREC', 'SCORE', and 'MATCH'. The table contains seven rows of data. A mouse cursor is pointing at the 'REFREC' column header.

	COMPREC	REFREC	SCORE	MATCH
1	8	2010	7.4404489	*
2	8	397	7.4404489	*
3	8	4230	7.4404489	*
4	8	2835	7.4404489	*
5	8	1615	7.4404489	*
6	8	1686	7.4404489	*
7	8	2816	-1.0554060	*

Estrutura

Esta opção mostra a estrutura (descrição dos campos) do arquivo aberto.

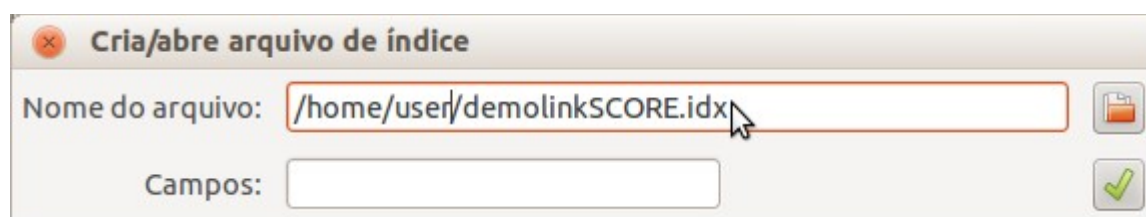
Reinicia

Os arquivos de dados podem ter um número muito grande de registros, mas a grade permite

a apresentação de cerca de 75 milhões de registros, no máximo. Caso o arquivo aberto tenha mais registros que este limite, o programa irá avisar, e só exibirá os registros abaixo do mesmo. A opção “Reinicia” permite que se escolha um novo ponto de origem para o arquivo, de modo a possibilitar a visualização em toda a sua extensão, ainda que por partes.

Índice

Este módulo permite que se abra um índice já existente, ou que se crie um novo. No segundo acaso, além de indicar o nome do arquivo, é necessário listar as variáveis que formarão a chave do índice, separadas por vírgulas. Caso se escolha essa opção e já exista um arquivo com o nome escolhido, o mesmo será sobrescrito (ver nota técnica sobre arquivos ao final do manual).



Movimenta

Este menu concentra as funções que transferem dados entre arquivos.

Anexa

Este módulo acrescenta os registros de um arquivo àquele já aberto. A estrutura dos dos não precisa ser a mesma, mas apenas os registros com mesmo nome e tipo serão copiados.

Exporta



Este módulo realiza a exportação do arquivo aberto para um arquivo texto. A primeira linha do texto será composta pela definição dos campos, e cada linha subsequente corresponde a um registro. Além dos campos do arquivo, o número de cada registro é também exportado. Os campos e definições de variáveis são separados pelo caractere definido como “separador de campos” (por padrão “|”), e cada parte da especificação da variável (nome, tipo, tamanho, decimais) é delimitada pelo caractere definido como “separador de definições” (por padrão a vírgula). Estes caracteres são livremente configuráveis, não podendo, contudo, serem iguais. Além disso, por razões óbvias, não se recomenda utilizar como separador caracteres comuns que poderiam ser parte de um campo, como números ou letras.

Importa

Importa arquivo texto

Separador de definições: , Separador de campos: |

Cabeçalho

☒ Primeira linha ☐ Arquivo separado ☐ Entrada manual

	Nome	Tipo	Tamanho	Decimais
1				
2				
3				
4				
5				

Registro corrente: 0

Adiciona Apaga Cancela OK

Este módulo realiza a operação inversa do anterior, importando um arquivo texto para o formato do programa. Além de definir os tipos de separador utilizados, é necessário informar onde estará a definição da estrutura do arquivo (o cabeçalho): se na primeira linha, como nos arquivos exportados pelo programa, se num arquivo à parte ou ainda se o usuário irá definir manualmente o cabeçalho para depois efetuar a importação. A estrutura será mostrada na grade, e os botões “Adiciona” e “Apaga” permitem adicionar linhas em branco ou apagar linhas da grade.

Ao final do processo o arquivo gerado na importação será aberto.

Importa dbf

Este módulo importa um arquivo no formato .dbf (também chamado xbase) para o formato do programa.

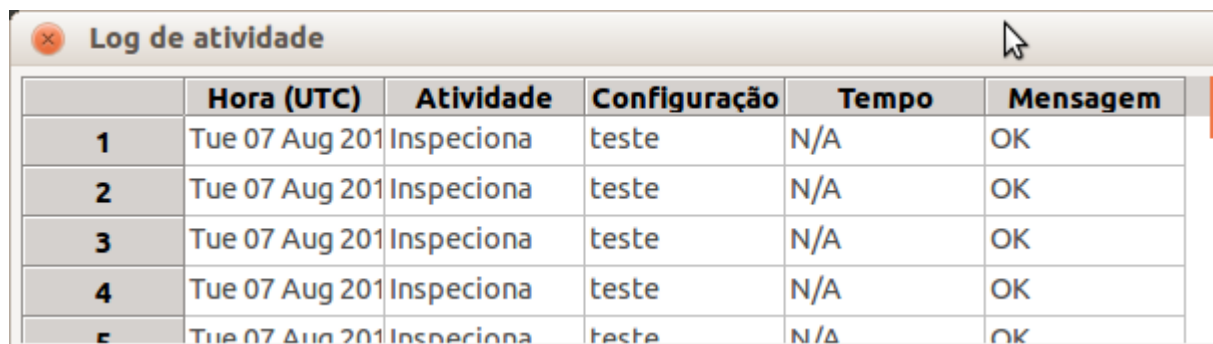
Ao final do processo o arquivo gerado na importação será aberto.

Sai

Esta opção fecha o módulo e quaisquer arquivos abertos.

Log Atividade

Este módulo simplesmente mostra o registro de atividade do programa, até o máximo de 1.000 operações mais recentes (as mais antigas vão sendo descartadas).

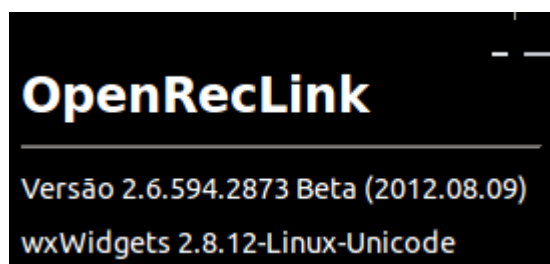


The screenshot shows a window titled "Log de atividade" with a close button in the top-left corner. The window contains a table with the following data:

	Hora (UTC)	Atividade	Configuração	Tempo	Mensagem
1	Tue 07 Aug 201	Inspeciona	teste	N/A	OK
2	Tue 07 Aug 201	Inspeciona	teste	N/A	OK
3	Tue 07 Aug 201	Inspeciona	teste	N/A	OK
4	Tue 07 Aug 201	Inspeciona	teste	N/A	OK
5	Tue 07 Aug 201	Inspeciona	teste	N/A	OK

Sobre

Este módulo apenas apresenta, no canto inferior direito da tela principal, a versão do programa e da biblioteca wxWidgets utilizada. Esta informação é importante para saber se a versão em uso é a mais atual. A janela se fecha automaticamente após alguns segundos de exibição.



Sai

Esta opção encerra a execução do programa.

Cópia

Este bloco contém as funções para a cópia seletiva (isto é, baseada em algum critério) de arquivos do programa.

Configura

Este assistente permite a definição do conjunto de opções que irá controlar o processo de cópia do arquivo.



The image shows a software window titled "Assistente" (Wizard) with a close button in the top-left corner. The main heading inside the window is "Cópia seletiva" (Selective Copy). Below this heading are four tabs: "Arq. entrada" (Input File), "Condições" (Conditions), "Campos" (Fields), and "Finaliza" (Finish). The "Arq. entrada" tab is currently selected. The main content area of the window features a large, empty text input field for specifying the source file name, and a small file icon button to the right of the input field. At the bottom of the window, there are two navigation buttons: "Volta" (Back) with a left-pointing arrow and "Segue" (Next) with a right-pointing arrow.

Primeiramente é necessário definir o nome do arquivo de origem, a partir do qual será feita a cópia.

Assistente

Cópia seletiva

Arq. entrada Condições Campos Finaliza

Campo	Op. lógico	Valor	Conector
SCORE	Igual a	NNN	ou
REFREC	Maior que	522	e
R_NAME	Menor ou i	ABCDEFGG	ou
R_NUMBER	Menor que	ADSDS	e
C_YEAR	Diferente		

< Volta Segue >

A seguir, é necessário definir as condições que, se atendidas, irão permitir que um registro do arquivo existente seja copiado para o novo arquivo. Cada condição é definida por um nome de campo, um operador lógico (maior que, maior ou igual a, igual a, menor ou igual a, menor que, diferente de) e um valor. Além disso, as condições podem ser agregadas utilizando-se conectores (“e”, “ou”). Na execução da cópia, cada condição é checada de cima para baixo, interrompendo a checagem no momento em que o conector é vazio. O registro só será copiado se a avaliação final das condições combinadas for “verdadeiro”.

Assistente

Cópia seletiva

Arq. entrada Condições Campos Finaliza

Selecione os campos desejados do arquivo mestre na caixa ao lado:

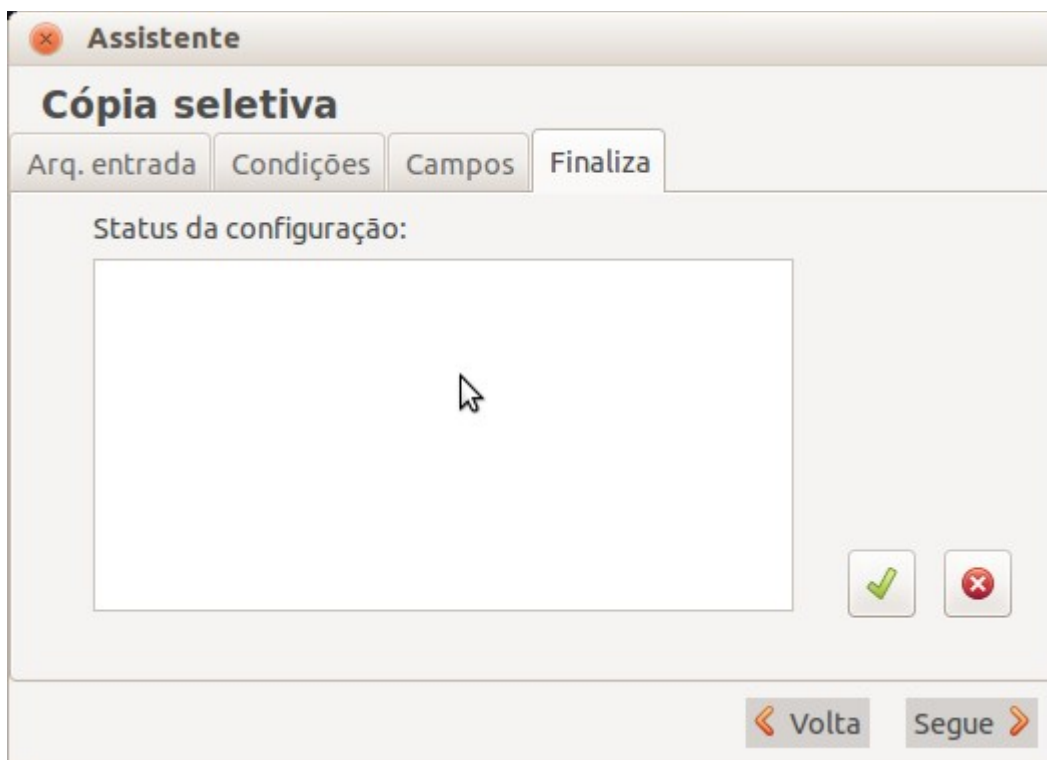
Marca todos campos Desmarca todos campos

- ☒ COMPREC
- ☐ REFREC
- ☐ SCORE
- ☐ R_NAME
- ☒ R_MOTHER
- ☒ R_BIRTH
- ☒ R_SEX
- ☐ R_NUMBER

< Volta Segue >

Por fim, é necessário definir-se que campos irão ser copiados para o novo arquivo. Somente

os campos checados serão copiados. Pelo menos um campo precisa ser selecionado. Os botões à esquerda e abaixo da lista de campos permitem marcar e desmarcar todos os campos em uma única operação.



Na última página do assistente eventuais erros serão mostrados. Só será permitido gravar a configuração se não houver nenhum erro.

Executa

Esta opção executa a cópia segundo a configuração produzida anteriormente. Deve-se indicar o arquivo de configuração e o nome do novo arquivo que será gerado, clicando-se no botão à esquerda e abaixo do nome selecionado para o último para dar início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



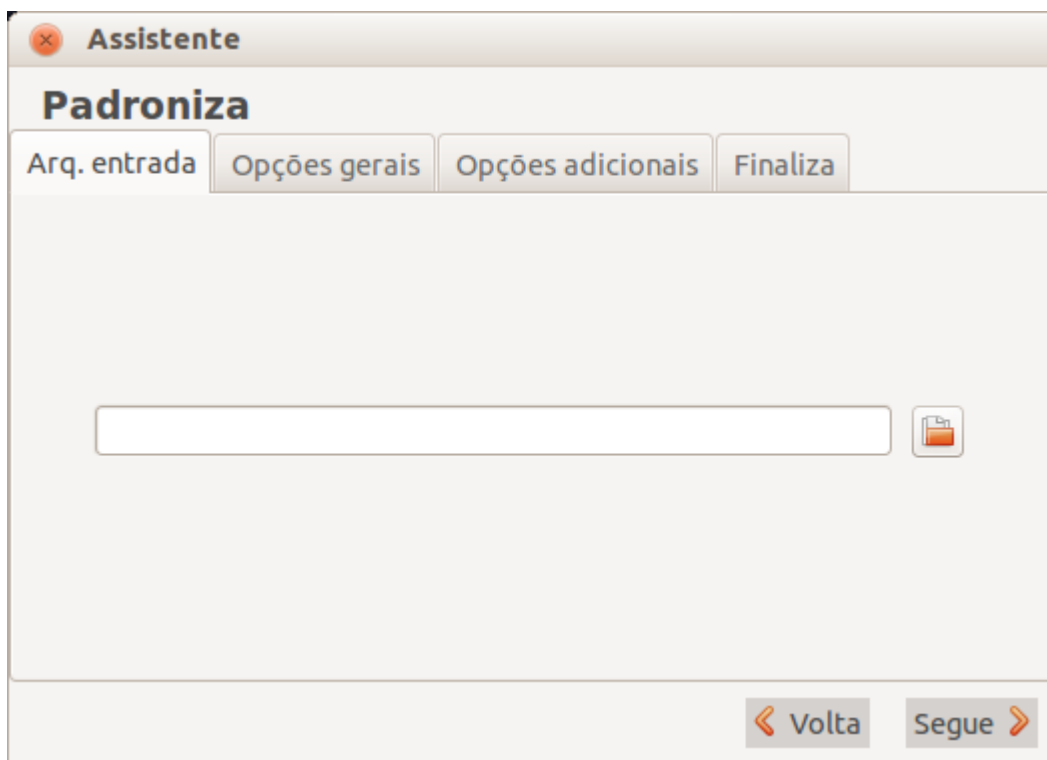
The image shows a software dialog box titled "Executa cópia seletiva". It contains two input fields: "Arq. configuração" and "Arq. saída", each with a folder selection icon to its right. Below these fields is a horizontal line. Under the line, on the left, is a gear icon. To the right of the gear icon is the text "Registro corrente: 0". Below this text is a long, empty, rounded rectangular box. At the bottom right of the dialog is a button with a right-pointing arrow icon.

Padroniza

Este bloco agrupa as funções para padronização de um arquivo, isto é, a configuração de seus campos em formatos adequados ao processo de relacionamento.

Configura

Este assistente configura as opções de padronização, especificando para cada campo de entrada o tipo de transformação e o nome do campo de saída.



The image shows a software window titled "Assistente" with a sub-header "Padroniza". Below the header is a tabbed interface with four tabs: "Arq. entrada", "Opções gerais", "Opções adicionais", and "Finaliza". The "Arq. entrada" tab is currently selected. Inside this tab, there is a large empty rectangular text input field. To the right of this field is a small icon representing a file. At the bottom right of the window, there are two buttons: "Volta" (with a left-pointing arrow) and "Segue" (with a right-pointing arrow).

O primeiro passo é definir um arquivo de entrada, para ser padronizado.

Assistente

Padroniza

Arq. entrada Opções gerais Opções adicionais Finaliza

	Entrada	Processo	Saída
1	NOME	CNOMES	NOME
2	NOME	FNOMEF	FNOMEF
3	NOME	FNOMEU	FNOMEU
4	NOME	FNOMEI	FNOMEI
5	NOME	FNOMEA	FNOMEA
6	NOME	PBLOCO	PBLOCO

Adiciona todos campos

Adiciona Remover Limpa

Volta Segue

A seguir deve-se criar a tabela de conversão; cada linha na tabela representa um campo de entrada, o processo de transformação e o campo de saída. Um mesmo campo de entrada pode aparecer várias vezes, mas não o de saída. Os botões abaixo da grade permitem adicionar, remover ou limpar o conteúdo da linha selecionada; o botão à esquerda da grade permite adicionar todos os campos de entrada à grade de uma só vez.

No processo de padronização todos os campos serão convertidos para campos caractere; campos no formato Data, Numérico ou Lógico serão automaticamente convertidos; campos no formato Memo serão ignorados; finalmente, campos do tipo Caractere serão convertidos de acordo com a especificação do usuário.

As opções de data visam a padronização de campos caractere que contenham informação sobre datas (campos data já são formatados corretamente por padrão).

As opções “Elimina pontuação”, “Nomes próprios” e “Subdivide cadeia”, tratam as cadeias de caracteres com ações que são progressivamente mais abrangentes. A opção “Elimina pontuação” retira todos os sinais de pontuação definidos pelo usuário. A opção “Nomes próprios” faz a mesma coisa que a opção “Elimina pontuação” e, adicionalmente, retira cadeias de caracteres definidas pelo usuário (ex. de, dos, das), elimina espaços duplos, elimina todos os acentos, elimina todos os dígitos e transforma todos os caracteres para caixa alta. A opção “Subdivide nome” faz as mesmas tarefas que a opção anterior e, adicionalmente, cria automaticamente seis campos com nomes padrão FNOMEF, FNOMEU, FNOMEI, FNOMEA, PBLOCO e UBLOCO. Estes campos armazenam, respectivamente, o primeiro nome, o último nome, as iniciais do meio, os apêndices (Jr., Filho, etc.) e o primeiro e último nomes formatados para blocagem (com pequenas modificações nas primeiras letras, para evitar problemas na utilização do código soundex para blocagem). Estes nomes podem ser mudados, e os campos omitidos. Sugerimos fortemente que ao menos o primeiro nome e o último, formatados para blocagem, sejam deixados, na medida em que são campos úteis para a blocagem (ver a rotina de Relacionamento).

Clicando na coluna do meio (Processo) para campos caractere, surge a tela abaixo, que permite selecionar a opção específica desejada, bem como visualizar o que a transformação irá gerar. As listas “Entrada” e “Saída” contém, respectivamente, um conjunto de valores existentes no

arquivo de entrada e os valores que apresentarão na saída uma vez aplicado o processo de conversão selecionado. Processos com saídas múltiplas terão as mesmas apresentadas em uma mesma linha separadas por vírgula.

Tipos de conversão

Tipos de conversão

- ☐ Data
- ☐ Nomes próprios
- ☐ Elimina pontuação
- ☐ Subcadeia
- ☒ Nenhuma (cópia simples)
- ☐ Subdivide nome
- ☐ Substitui valores

Tipos de data

- ☒ DD/MM/AA
- ☐ DDMMAA
- ☐ DD/MM/AAAA
- ☐ DDMMAAAA
- ☐ MM/DD/AA
- ☐ MMDDAA
- ☐ MM/DD/AAAA
- ☐ MMDDAAAA
- ☐ AA/MM/DD
- ☐ AAMMDD
- ☐ AAAA/MM/DD
- ☐ AAAAMMDD

Substituições

	Troca	Por
1		
2		
3		
4		
5		
6		
7		
8		

Refaz

Início

1

Caracteres

1

Entrada

19011918
29031933
28101925
25121925
10081927
11051927
10031942
03011934
24081920
20071940

Saída

14021930
14031926
18111929
25061916
21031956
17081921
08081923

Cancel OK

A opção “Subcadeia” habilita os campos “Início” e “Caracteres”, que configuram, respectivamente, o caractere onde a cópia se inicia e quantos caracteres serão copiados para o novo campo (por exemplo, se o campo tiver o conteúdo “GUARDA-CHUVA”, e for selecionado início 8 e 5 caracteres, o resultado será “CHUVA”).

A opção “Substitui valores” habilita a tabela de substituições, ao centro da janela. Para cada valor encontrado numa linha à esquerda, o novo campo será preenchido com o valor da direita (por exemplo, se a linha contiver “1” e “Masculino”, toda vez que o campo de entrada selecionado tiver o valor “1” o campo de saída irá conter “Masculino”). Dois caracteres, se colocados à esquerda, terão interpretação especial: “.” significa campo vazio, e “*” significa qualquer conteúdo além dos anteriormente especificados (coloque este caractere ao final da lista de valores para substituição, ou os resultados não serão os esperados). O botão “Refaz” apresenta as transformações resultantes na janela respectiva.

The screenshot shows a window titled 'Assistente' with a sub-header 'Padroniza'. It has four tabs: 'Arq. entrada', 'Opções gerais', 'Opções adicionais' (which is active), and 'Finaliza'. Under the 'Opções adicionais' tab, there are three text input fields. The first is labeled 'Exclusões:' and contains the text 'DE,DO,DA,DOS,DAS'. The second is labeled 'Pontuação:' and contains the text '()/'.,:;-'. The third is labeled 'Apêndices:' and contains the text 'JUNIOR JR FILHO FO NETO SOBRINHO'. At the bottom right of the window, there are two buttons: '< Volta' and 'Segue >'. A mouse cursor is pointing at the 'Segue >' button.

Neste ponto é possível indicar que partículas serão eliminadas na padronização de nomes próprios, que caracteres são sinais de pontuação (também eliminados na padronização de nomes) e que tipo de nome é considerado um apêndice.

The screenshot shows the same 'Assistente Padroniza' window, but now the 'Finaliza' tab is active. The 'Status da configuração:' label is at the top of the main area, above a large empty rectangular box. To the right of this box are two small square buttons: one with a green checkmark and one with a red 'X'. At the bottom right of the window, there are two buttons: '< Volta' and 'Segue >'. A mouse cursor is pointing at the 'Finaliza' tab.

Na última página do assistente eventuais erros serão mostrados. Só será permitido gravar a configuração se não houver nenhum erro.

Executa

Esta opção executa a cópia segundo a configuração produzida anteriormente. Deve-se indicar o arquivo de configuração e o nome do novo arquivo que será gerado, clicando-se no botão à esquerda e abaixo do nome selecionado para o último para dar início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



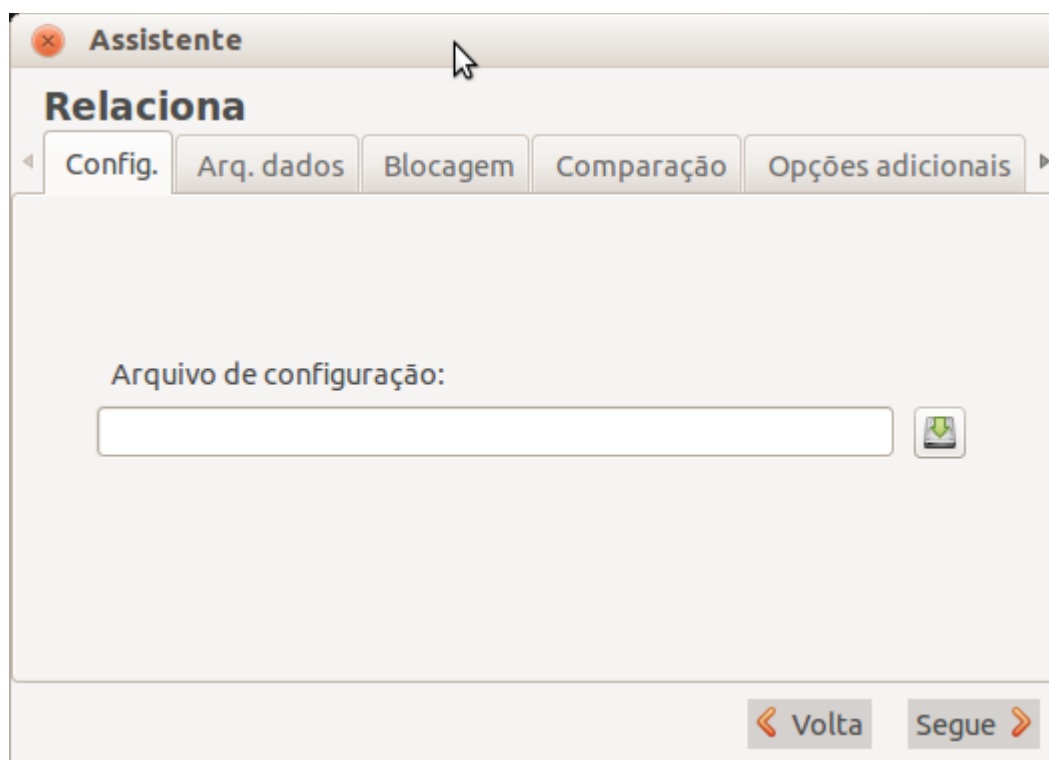
The image shows a software window titled "Executa padronização" (Execute standardization). It contains two input fields for file selection: "Arq. configuração" (Configuration file) and "Arq. saída" (Output file). Each field has a folder icon button to its right. Below these fields is a horizontal line. Under the line, on the left, is a gear icon. To its right is the text "Registro corrente: 0" (Current record: 0). Below this text is a long, empty, rounded rectangular box. At the bottom right of the window is a button with a red arrow pointing left.

Relaciona

Este bloco agrupa as funções para o relacionamento probabilístico de dois arquivos, bem como várias funções acessórias.

Configura

Este assistente permite configurar os parâmetros do relacionamento de registros.



Inicialmente deve-se selecionar um nome para o arquivo de configuração.

Assistente

Relaciona

◀ Config. Arq. dados Blocagem Comparação Opções adicionais ▶

Arquivo de comparação (será lido integralmente)

Arquivo de referência (será lido por blocos):

< Volta Segue >

A seguir deve-se selecionar os arquivos que serão selecionados, conforme a definição acima.

Assistente

Relaciona

◀ Config. Arq. dados Blocagem Comparação Opções adicionais ▶

	Comparação	Referência
1	PBLOCO	PBLOCO
2	UBLOCO	UBLOCO
3	SEX	SEX

Nome do índice:

Adiciona Remover Limpa

< Volta Segue >

Neste ponto definem-se as opções de blocagem, indicando os campos em cada arquivo que permitirão a formação de blocos. Os nomes de campos não necessitam serem iguais, mas devem ser correspondentes (mesmo tipo de dado e conteúdo). Para o processo de blocagem será criado um índice para o arquivo de referência; a caixa de texto à esquerda permite selecionar o nome do mesmo (o nome final será transformado para caixa alta).

As teclas abaixo da grade permitem adicionar, remover ou limpar a linha selecionada. Os

botões à direita da grade permitem mudar a ordem das linhas.

Assistente

Relaciona

◀ Config. Arq. dados Blocagem **Comparação** Opções adicionais ▶

	Comp.	Tipo	Corr.	Inc.	Lím.	Ref.
1	NAME	APROX	95	0,16	85	NAME
2	BIRTH	CARACTE	90	1,4	65	BIRTH
3	FNOMEF	APROX	95	5	85	FNOMEF
4	FNOMEU	APROX	95	5	85	FNOMEU

Adiciona Remover Limpa

◀ Volta Segue ▶

Nesta grade é possível selecionar os campos e processos de comparação, bem como os respectivos parâmetros (m, u e limiar). Os botões têm função homóloga aos da página anterior do assistente.

Assistente

Relaciona

◀ Arq. dados Blocagem Comparação **Opções adicionais** Finaliza ▶

Limites de escores possíveis com a configuração corrente:

Máximo: 23,716

Mínimo: -16,1171

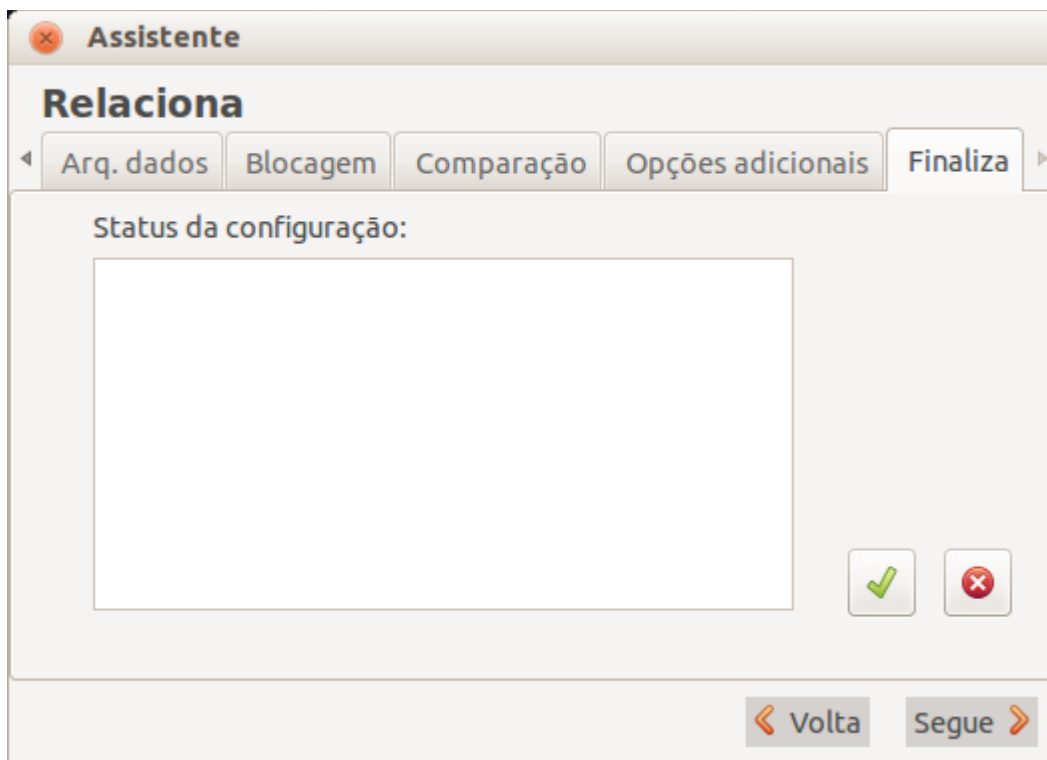
☒ Ignora escore baixo:

Escore mínimo:

-2

◀ Volta Segue ▶

Aqui é possível determinar um valor mínimo de score para que o par gerado seja gravado (sobre pares, ver o item “Executa”, mais à frente).



Na última página do assistente eventuais erros serão mostrados. Só será permitido gravar a configuração se não houver nenhum erro.

Gera freq

Neste módulo é possível gerar uma tabela de frequência para os valores existentes num determinado campo, como forma de determinar a probabilidade de designação errônea de um par como verdadeiro. Deve-se indicar o arquivo de configuração e que pares de campos serão tabulados.

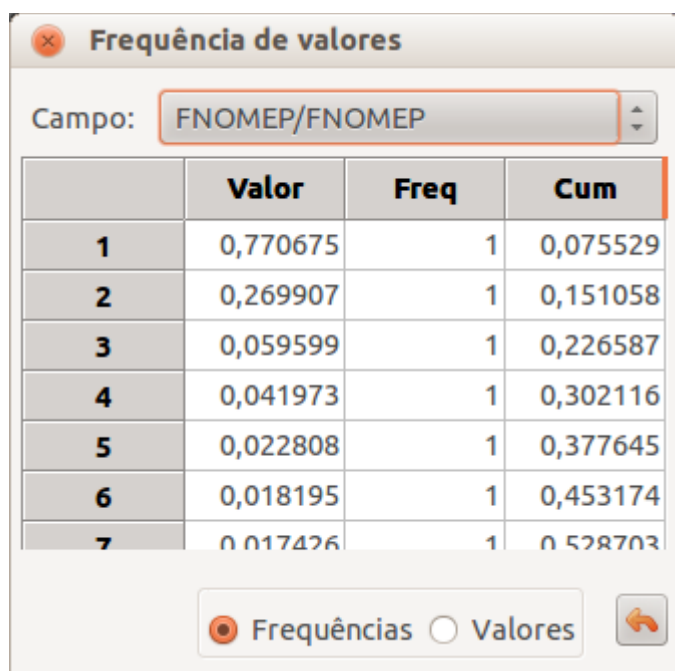


The image shows a software dialog box titled "Calcula frequência de valores em campos". It contains the following elements:

- A label "Arq. configuração:" followed by a text input field and a file selection icon (a folder with a document).
- A label "Campos para determinar freq." followed by a large empty rectangular box for selecting fields.
- A horizontal separator line.
- A small gear icon on the left and the text "Registro corrente: 0" in the center.
- A long, empty horizontal bar below the separator.
- A bottom section with a right-pointing arrow icon.

Analisa freq

Neste módulo é possível visualizar os resultados da tabulação realizadas no módulo anterior. É possível visualizar-se com que frequência determinadas probabilidades foram identificadas no par de arquivos, conforme a figura abaixo:



Frequência de valores

Campo: FNOME/FNOME

	Valor	Freq	Cum
1	0,770675	1	0,075529
2	0,269907	1	0,151058
3	0,059599	1	0,226587
4	0,041973	1	0,302116
5	0,022808	1	0,377645
6	0,018195	1	0,453174
7	0,017426	1	0,528703

☒ Frequências ☐ Valores

É possível ainda observar a frequência de valores específicos nos arquivos combinados, conforme a figura abaixo:



Frequência de valores

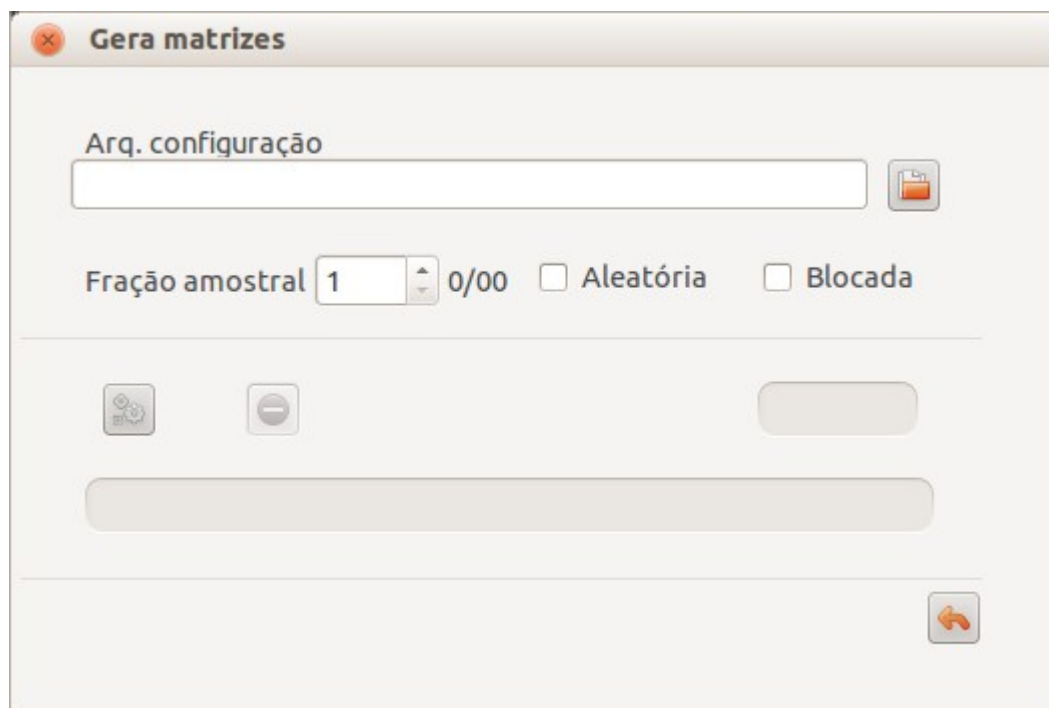
Campo: FNOME/FNOME

	Valor	Freq	Cum
1	MARIA	0,770675	0,770675
2	JOSE	0,269907	1,040582
3	ANTONIO	0,059599	1,100181
4	JOAO	0,041973	1,142154
5	MANOEL	0,022808	1,164962
6	LUIZ	0,018195	1,183157
7	JORGE	0,017426	1,200583

☐ Frequências ☒ Valores

Gera matrizes

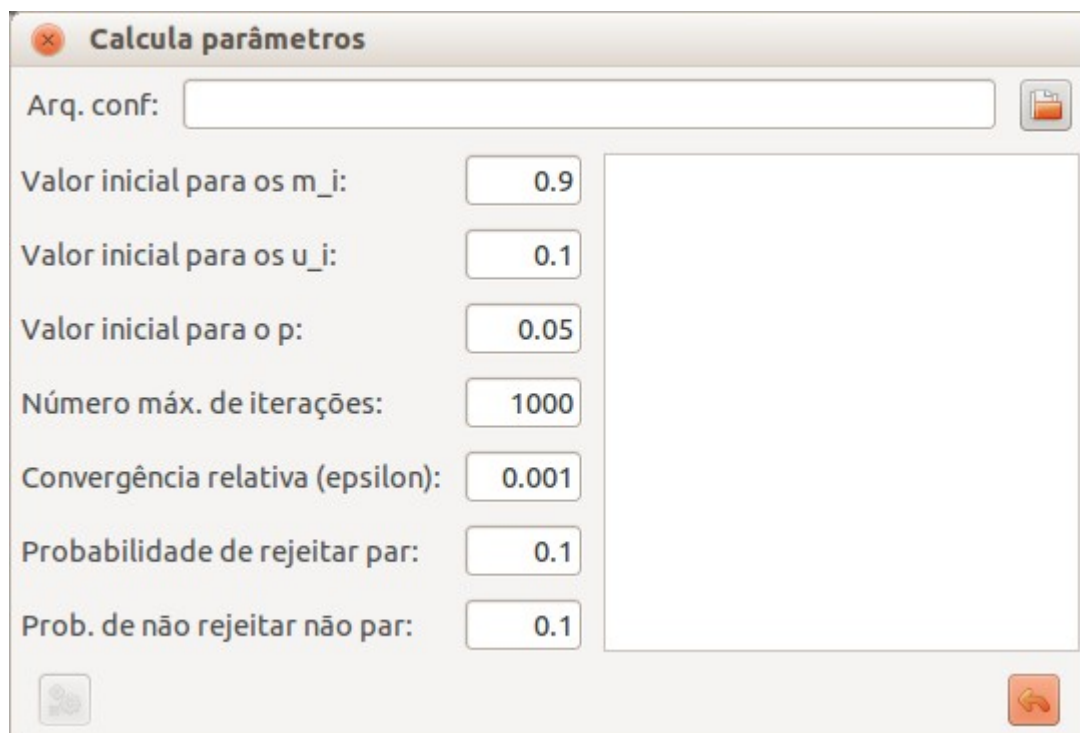
Este módulo gera as matrizes para utilização no módulo de cálculo de parâmetros (ver mais adiante). São gerados dois tipos de matrizes: aleatória (comparando aleatoriamente registros dos dois arquivos) e blocada (seguindo o processo de blocagem utilizado no relacionamento propriamente dito). A rotina de cálculo necessita das duas, mas elas podem ser geradas separadamente. Caso não haja nenhum par na matriz aleatória pode-se tentar executar a rotina novamente com uma fração amostral maior (ver abaixo).



Deve-se indicar o arquivo de configuração, e cada checkbox abaixo deste seleciona se um tipo de matriz será criada ou não. Para a matriz aleatória é necessário selecionar-se o tamanho da fração amostral que será utilizada. Clicando-se no botão à esquerda e abaixo da fração amostral inicia o processo. O botão à direita deste interrompe a execução da rotina. O botão no canto inferior direito fecha a janela, encerrando a operação.

Calcula

Este módulo permite otimizar os parâmetros do arquivo de configuração, a partir de parâmetros iniciais (ver abaixo) e das matrizes geradas no módulo anterior. Ao final do processo, é possível transferir automaticamente os valores calculados para a configuração escolhida.



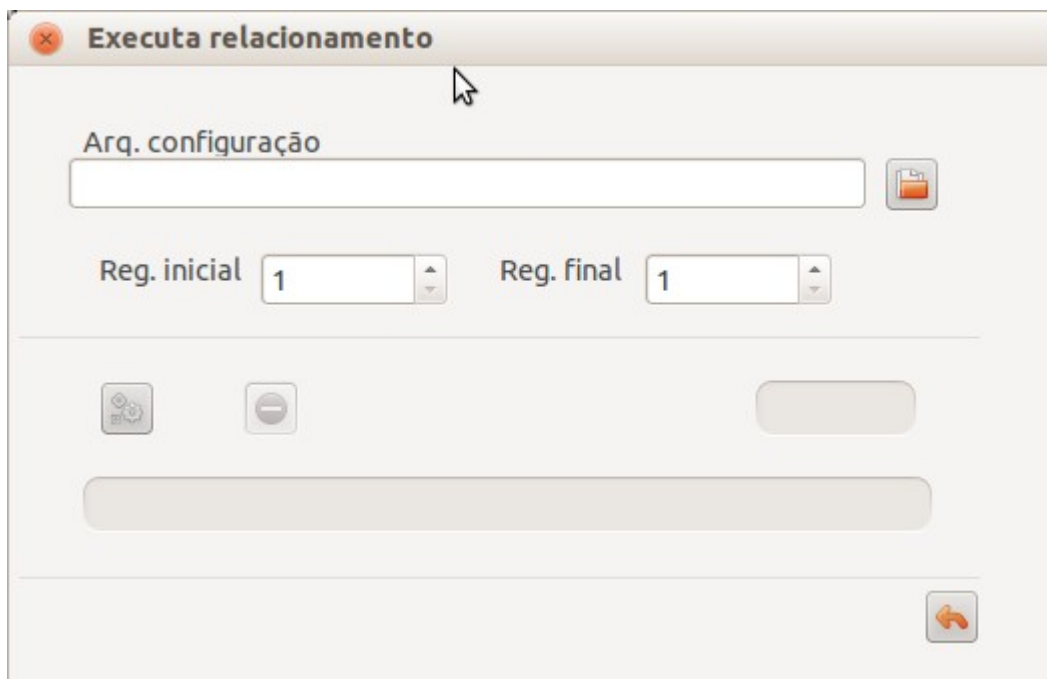
The image shows a software window titled "Calcula parâmetros". It contains a file selection field labeled "Arq. conf:" with a folder icon. Below this are seven input fields for numerical parameters, each with a label and a text box containing a value. To the right of these fields is a large empty rectangular area. At the bottom left is a small icon of a gear, and at the bottom right is a red button with a curved arrow pointing right.

Label	Value
Valor inicial para os m_i :	0.9
Valor inicial para os u_i :	0.1
Valor inicial para o p :	0.05
Número máx. de iterações:	1000
Convergência relativa (epsilon):	0.001
Probabilidade de rejeitar par:	0.1
Prob. de não rejeitar não par:	0.1

NOTA: Os números devem ser especificados segundo a configuração do computador; apesar da imagem acima, num computador configurado para a linha portuguesa os números decimais devem ser separados por **vírgulas** e não pontos. Não atentar para este ponto levará a erros.

Executa

Este módulo executa o processo de relacionamento tal como definido no módulo de configuração. Deve-se indicar o arquivo de configuração, e os registros inicial e final (do arquivo de referência) para processamento. Clicando-se no botão à esquerda e abaixo do registro inicial dá início ao processo. O botão à direita deste interrompe a execução da rotina. O botão no canto inferior direito fecha a janela, encerrando a operação.



Este módulo gera um arquivo denominado **arquivo de pares**, que terá o mesmo nome do arquivo de configuração seguido de "SCORE", com a extensão de arquivo de dados do programa (ver nota técnica). Cada registro deste arquivo contém o número do registro do arquivo de referência, o número do registro do arquivo de comparação, o score final da comparação das variáveis dos dois e um campo de identificação do par gerado (ver o item "Marca/Desmarca", mais à frente). Caso se tenha utilizado a opção de score mínimo, só serão gravados os pares cujo score seja superior ao limite designado.

Reduz pares

Este módulo executa cópia seletiva de um arquivo de pares, deixando de fora aqueles que não atendem ao critério de seleção. Pode-se optar por copiar apenas os registros marcados como pares (ver o item seguinte), ou aqueles com score maior ou igual a um valor escolhido.



The image shows a software window titled "Reduz pares" with a close button (X) in the top-left corner. The window contains the following elements:

- A text field labeled "Arq. configuração" with a file selection icon (folder with a document) to its right.
- A section titled "Tipo de seleção" containing two radio buttons: "Pares marcados" (unselected) and "Valor mínimo" (selected). To the right of the "Valor mínimo" radio is a text input field containing the number "0".
- Below the selection options, there are two icons: a gear icon and a minus sign icon.
- To the right of these icons is the text "Registro corrente: 0".
- Below this text is a horizontal progress bar.
- In the bottom-right corner, there is a button with a right-pointing arrow.

Deve-se indicar o arquivo de configuração, e o tipo de cópia (pares marcados ou valor mínimo) para processamento; no segundo caso, deve-se digitar ainda o valor mínimo de score desejado. Clicando-se no botão à esquerda e abaixo do item “pares marcados” dá início ao processo. O botão à direita deste interrompe a execução da rotina. O botão no canto inferior direito fecha a janela, encerrando a operação.

Marca/Desmarca

Este módulo permite alterar em bloco o campo de marcação e um arquivo de pares

Deve-se indicar o arquivo de configuração, e o tipo de operação (marcar ou desmarcar); no segundo caso, deve-se digitar ainda os valores de score desejado para marcação. Clicando-se no botão à esquerda e abaixo da opção de operação dá início ao processo. O botão à direita deste interrompe a execução da rotina. O botão no canto inferior direito fecha a janela, encerrando a operação.



The screenshot shows a window titled "Marca/Desmarca arquivo de pares". It contains a text field labeled "Arq. configuração" with a file selection icon to its right. Below this is a section labeled "Opção" with two radio buttons: "Desmarca" and "Marca". The "Marca" option is selected. To the right of the radio buttons are two text input fields labeled "Min. par" and "Max. n/par", both containing the value "0". Below these fields is a horizontal bar with a minus sign icon on the left and the text "Registro corrente: 0". At the bottom right of the window is a button with a right-pointing arrow icon.

Um campo pode ser marcado como “Ignorado” (isto é, não há informação sobre par ou não; este é o valor padrão ao final do processo de relacionamento, e a opção “Desmarca” restaura este valor), “Par”, “Dúvida” ou “Não par”. No caso da opção de marcação, registros com scores abaixo do valor designado como “Max n/par” serão marcados como “Não par”; acima do valor designado como “Min. par” serão marcados como “Par”, e aqueles com valores entre os dois serão marcados como “Dúvida”.

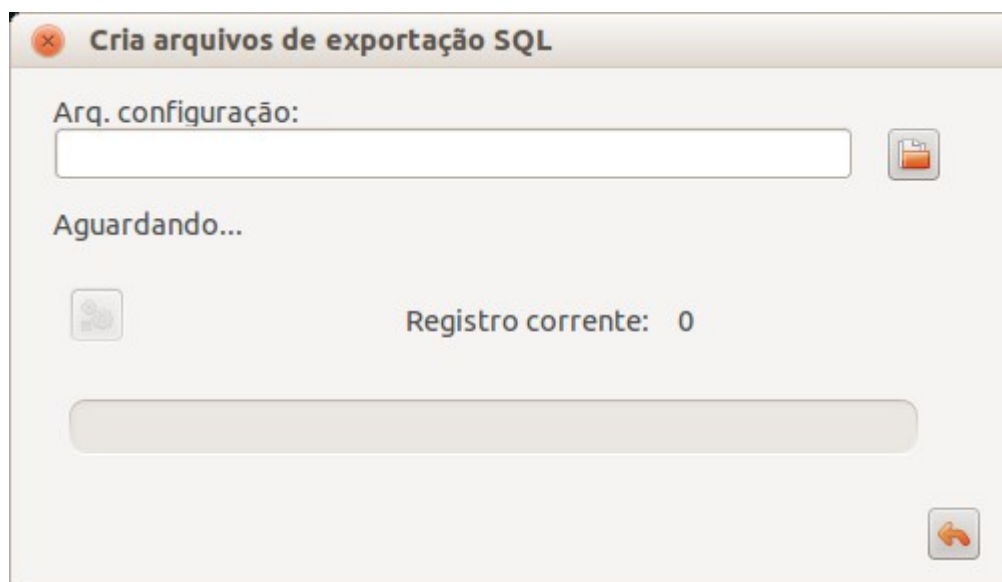
Exporta SQL

Este módulo cria uma série de arquivos de texto contendo:

- arquivo de referência;
- arquivo de comparação;
- arquivo de pares;
- comandos SQL para criar a estrutura e importar as tabelas anteriores;
- comando SQL para criar um *join* lateral das tabelas anteriores.

Os nomes dos arquivos criados serão os mesmos dos arquivos originais, trocando-se as extensões para “.csv”.

Deve-se indicar o arquivo de configuração. Clicando-se no botão à esquerda e abaixo do nome selecionado para o último dá início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



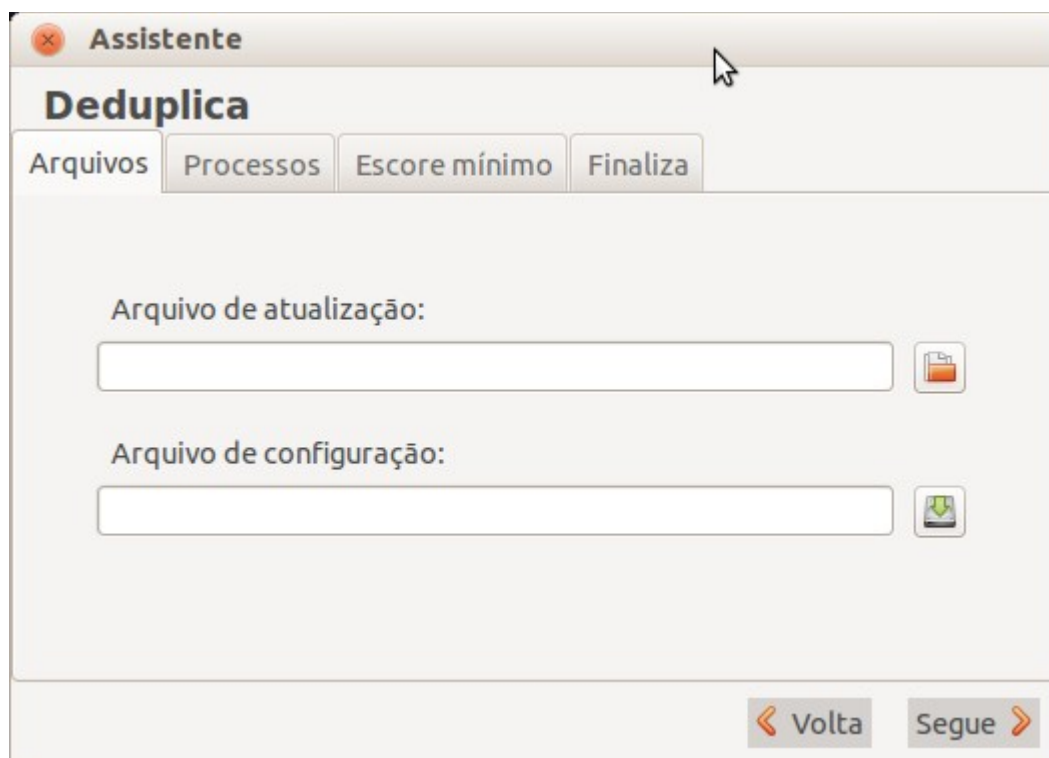
Este módulo foi criado como alternativa para examinar arquivos muito grandes (com dezenas de milhões de registros cada), com mais rapidez que na rotina de exame do módulo de combinação (ver abaixo). Os comandos e arquivos criados foram testados com SQL, mas são arquivos texto que podem ser facilmente editados para adaptação a outros SGBD.

Deduplica

Este módulo permite identificar registros duplicados num arquivo com base no mesmo método utilizado no relacionamento probabilístico. O arquivo original será copiado para um novo, contendo dois campos para armazenar o score atingido e um identificador único. Registros que atingem um score superior a um valor determinado terão um mesmo indicador único.

Configura

Este assistente permite configurar as opções para a deduplicação. Elas são basicamente homólogas ao processo de relacionamento, e por este motivo apenas serão descritos naquilo em que diferem.



A imagem mostra a interface de um assistente de deduplicação. A janela tem o título "Assistente" e o subtítulo "Deduplica". Há quatro abas: "Arquivos", "Processos", "Escore mínimo" e "Finaliza". O campo "Arquivo de atualização:" tem um ícone de pasta ao lado. O campo "Arquivo de configuração:" tem um ícone de download ao lado. Na base da janela, há dois botões: "Volta" com uma seta para trás e "Segue" com uma seta para frente.

Inicialmente é preciso escolher o arquivo que será deduplicado e um nome para o arquivo de configuração.

Assistente

Deduplica

Arquivos
Processos
Escore mínimo
Finaliza

	Campo	Bloco	Comp.	Corr.	Inc.	Lim.
1	PBLOCO	<input checked="" type="checkbox"/>				
2	UBLOCO	<input checked="" type="checkbox"/>				
3	SEX	<input checked="" type="checkbox"/>				
4	YEAR	<input checked="" type="checkbox"/>				
5	NAME	<input type="checkbox"/>	APROX	95	2	85
6	MOTHER	<input type="checkbox"/>	APROX	95	2	85

Adiciona
Remover
Limpa

< Volta
Segue >

Esta página do assistente agrega as funções de blocagem e comparação. Clicando-se na coluna “Bloco”, um campo pode ser selecionado para compor ou não o critério de blocagem; caso esta coluna esteja ativa, não é possível entrar dados nas demais. As outras colunas são homólogas ao relacionamento; a única diferença é que só é necessário identificar um campo, já que o arquivo será comparado com ele mesmo.

Assistente

Deduplica

Arquivos
Processos
Escore mínimo
Finaliza

Limites de escores possíveis
com a configuração corrente:

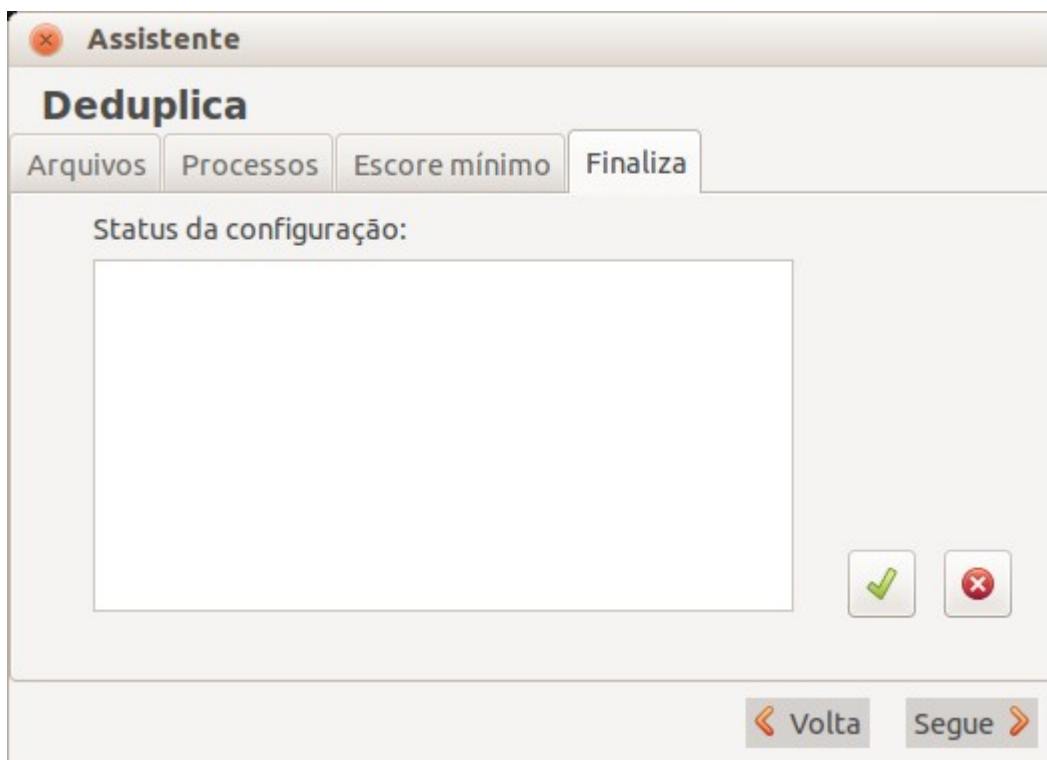
Máximo: 19,4796

Mínimo: -15,0814

Escore mínimo:

< Volta
Segue >

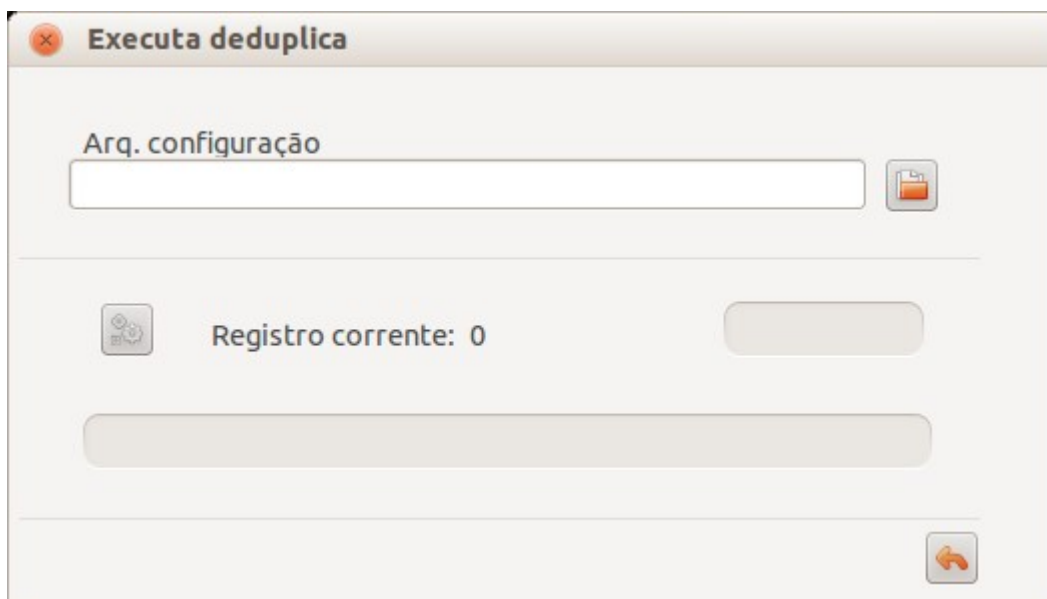
O escore mínimo irá definir que registros serão marcados com o mesmo identificador.



Na última página do assistente eventuais erros serão mostrados. Só será permitido gravar a configuração se não houver nenhum erro.

Executa

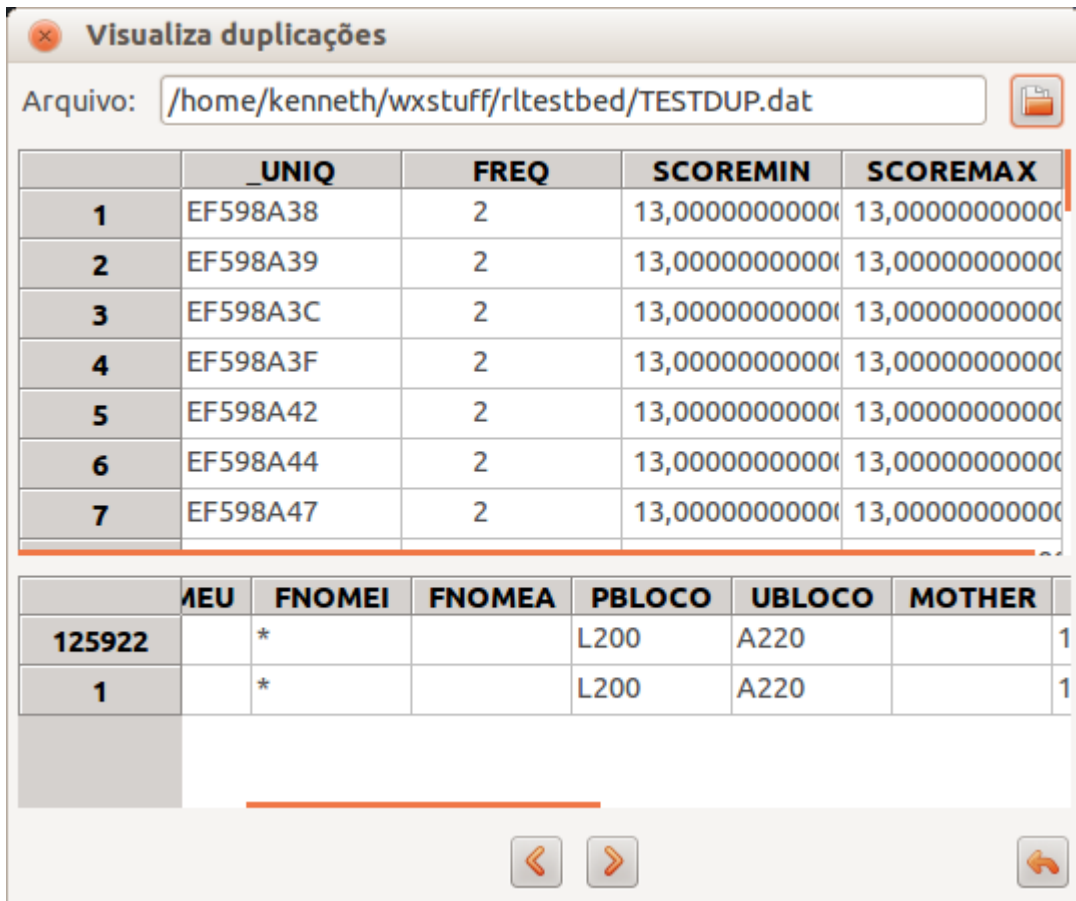
Este módulo executa o processo de deduplicação tal como definido no módulo de configuração. Deve-se indicar o arquivo de configuração para processamento. Clicando-se no botão à esquerda e abaixo do nome do arquivo dá início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



Examina

Este módulo permite examinar o arquivo gerado pelo processo de deduplicação. Na parte superior da janela são mostrados os indicadores únicos, sua frequência de ocorrência no arquivo e os scores mínimo e máximo associados àquele identificador. Na parte inferior os registros referentes a um determinado identificador são mostrados.

Para possibilitar essa visualização, um arquivo de sumário (com o mesmo nome do arquivo, contendo “SUM” acrescentado ao final deste), bem como outro índice, são criados.



The screenshot shows a window titled "Visualiza duplicações" with a file path input field set to "/home/kenneth/wxstuff/rltestbed/TESTDUP.dat". Below the input field are two tables. The first table is a summary of unique identifiers, and the second table shows detailed records for a selected identifier.

	_UNIQ	FREQ	SCOREMIN	SCOREMAX
1	EF598A38	2	13,000000000000	13,000000000000
2	EF598A39	2	13,000000000000	13,000000000000
3	EF598A3C	2	13,000000000000	13,000000000000
4	EF598A3F	2	13,000000000000	13,000000000000
5	EF598A42	2	13,000000000000	13,000000000000
6	EF598A44	2	13,000000000000	13,000000000000
7	EF598A47	2	13,000000000000	13,000000000000

	MEU	FNOMEI	FNOMEA	PBLOCO	UBLOCO	MOTHER	
125922		*		L200	A220		1
1		*		L200	A220		1

At the bottom of the window, there are three buttons: a left arrow, a right arrow, and a close button (X).

Os dois botões ao centro na parte inferior da janela deslocam a grade de registros para o identificador anterior ou seguinte, respectivamente; o botão no canto inferior direito fecha a janela.

Combina

Este bloco agrupa as funções de geração de arquivo combinado, a partir de processo de relacionamento executado anteriormente.

Examina

Este módulo cria um arquivo virtual associando os arquivos de comparação e referência, com base no arquivo de pares. Para iniciar seu funcionamento, é necessário selecionar o arquivo de configuração.

Ao iniciar a operação, o módulo irá gerar um índice para o arquivo de pares, passando depois a tabular a frequência dos scores; o programa avisa caso estes arquivos já existam, dando ao usuário a opção de utilizá-los ou não.

A grade à esquerda mostra os campos do arquivo de referência, e a da direita, os do arquivo de comparação. Os botões com setas para cima e para baixo, à esquerda da primeira grade e à direita da segunda, permitem que se altere a ordem dos campos na grade respectiva.

A grade inferior mostra a distribuição dos scores no arquivo de pares. O grupo de seleção à direita mostra a marcação do registro corrente (ver o item “Marca/Desmarca”, anteriormente neste manual), e permite alterá-la, bastando clicar na opção desejada. Os campos “Score + min” e “Score – max”, bem como os botões “Marca todos” e “Limpa todos” têm a mesma função da rotina “Marca/Desmarca”. Os botões “Fim”, “Avança”, “Volta” e “Início” permitem o deslocamento dentro de um mesmo bloco de scores; o bloco desejado é selecionado clicando-se na linha correspondente da tabela de frequência. A opção “passo” controla o deslocamento dos botões “Avança” e “Volta”, determinando o número (entre 1 e 10) de registros “pulados” em cada caso.

Executa

Este módulo gera um novo arquivo, combinando os arquivos de referência e comparação, de acordo com a marcação efetuada no arquivo de pares. Deve-se indicar o arquivo de configuração para processamento, e o nome do novo arquivo resultante. Caso se tenha adotado o processo de múltiplos passos, é possível ainda gerar novos arquivos de referência e comparação, contendo apenas registros ainda não pareados. Caso esta opção seja selecionada, clicando-se no checkbox correspondente, é necessário indicar os nomes dos novos arquivos. Clicando-se no botão à esquerda e abaixo destes dá-se início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



Executa combinação

Arq. configuração de relacionamento

Arq. combinado

Novo arq. referência

Novo arq. comparação

☐ Gera novos arqs.

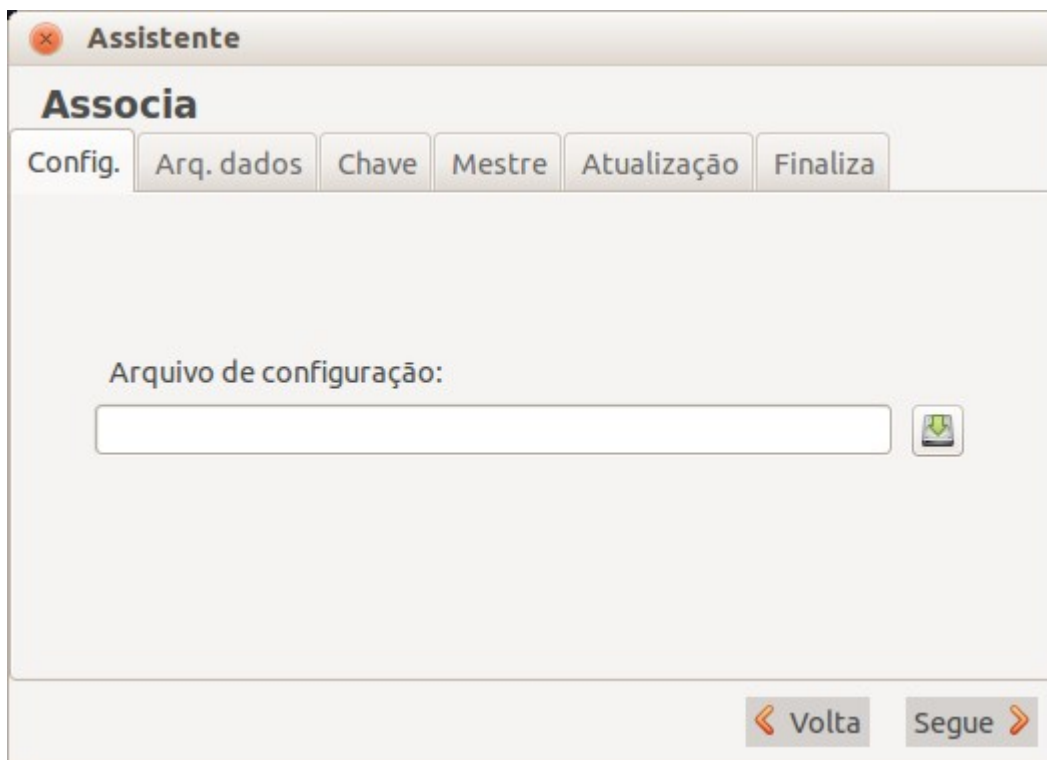
 Registro corrente: 0

Associa

Este bloco agrupa as funções para a associação de arquivos. Denominamos de “associação” o processo de junção de dois arquivos com base num campo-chave comum, compartilhado.

Configura

Este módulo configura as opções para a associação de dois arquivos.



The screenshot shows a software window titled "Assistente" (Wizard) with a close button. Below the title bar, the word "Associa" is displayed. A horizontal tab bar contains six tabs: "Config." (selected), "Arq. dados", "Chave", "Mestre", "Atualização", and "Finaliza". The main area of the window is labeled "Arquivo de configuração:" and contains a text input field. To the right of the input field is a small icon of a document with a green arrow pointing down. At the bottom right of the window, there are two buttons: "< Volta" and "Segue >".

Inicialmente deve-se selecionar um nome para o arquivo de configuração.

The screenshot shows a window titled "Assistente" with a close button. Below the title bar is a tabbed interface with the following tabs: "Config.", "Arq. dados" (which is the active tab), "Chave", "Mestre", "Atualização", and "Finaliza". The main area of the window contains two labels with corresponding text input fields and file selection icons:

- Label: "Arquivo mestre (será indexado)"
- Input field: [Empty text box]
- Icon: [File selection icon]
- Label: "Arquivo de atualização (será lido por inteiro):"
- Input field: [Empty text box]
- Icon: [File selection icon]

At the bottom right of the window, there are two buttons: "< Volta" and "Segue >".

A seguir, selecionam-se os nomes dos dois arquivos que serão associados.

The screenshot shows the same "Assistente" window, but now the "Chave" tab is active. The main area contains two labels with corresponding dropdown menus:

- Label: "Campo-chave do arquivo mestre:"
- Dropdown menu: [Empty dropdown box]
- Label: "Campo-chave do arquivo de atualização:"
- Dropdown menu: [Empty dropdown box]

At the bottom right of the window, there are two buttons: "< Volta" and "Segue >".

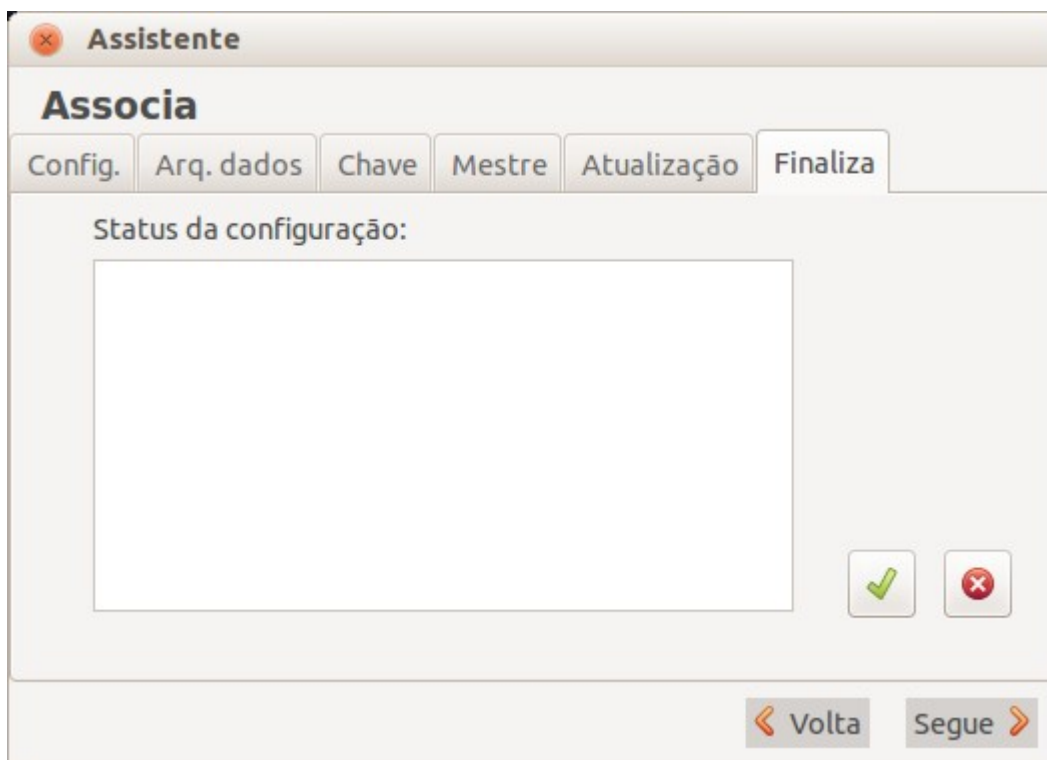
Aqui devem ser indicados os campos-chave dos dois arquivos; o pseudocampo “@RECNO” contém a informação do número de registro; isto permite que se reassocie arquivos previamente separados, desde que a informação do número de registro tenha sido preservada.

The screenshot shows a window titled 'Assistente' with a close button. Below the title bar is a tabbed interface with the following tabs: 'Config.', 'Arq. dados', 'Chave', 'Mestre', 'Atualização', and 'Finaliza'. The 'Mestre' tab is currently selected. The main area of the window contains the text 'Selecione os campos desejados do arquivo mestre na caixa ao lado:' followed by a large, empty rectangular box for selection. At the bottom right, there are two buttons: '< Volta' and 'Segue >'.

Aqui seleccionam-se os campos do arquivo mestre que serão incluídos no arquivo combinado.

This screenshot is similar to the previous one, showing the 'Assistente' window with the 'Associa' tab selected. In this step, the 'Atualização' tab is selected. The main area contains the text 'Selecione os campos desejados do arquivo de atualização na caixa ao lado:' followed by a large, empty rectangular box for selection. The bottom buttons '< Volta' and 'Segue >' are also present.

Aqui seleccionam-se os campos do arquivo de atualização que serão incluídos no arquivo combinado.



Na última página do assistente eventuais erros serão mostrados. Só será permitido gravar a configuração se não houver nenhum erro.

Executa

Este módulo executa o processo de associação segundo a configuração definida anteriormente. Deve-se indicar o arquivo de configuração para processamento, e o nome do arquivo resultante. Clicando-se no botão à esquerda e abaixo deste último dá-se início ao processo. O botão no canto inferior direito fecha a janela, encerrando a operação.



The image shows a software window titled "Executa associação" (Execute association). It contains two input fields: "Arq. configuração" (Configuration file) and "Arq. saída" (Output file). Each field has a text box and a folder icon button to its right. Below these fields is a horizontal line. Under the line, on the left, is a gear icon. To the right of the gear icon is the text "Registro corrente: 0". Below this text is a long, empty, rounded rectangular box. At the bottom right of the window is a button with a red arrow pointing left.

Notas técnicas

Os arquivos de dados são arquivos binários compostos por um cabeçalho contendo as definições de variáveis, seguidos por registros de tamanho fixo com os campos ajustados à esquerda no seu tamanho definido. Todos os dados são armazenados como sequências de caracteres de um byte.

A extensão do arquivo de dados depende da arquitetura do sistema operacional; em sistemas de 32 bits a extensão será “.d32”, e nos de 64 bits, “.dat”. Cada campo pode ter no máximo 1kByte de extensão, e o registro no máximo 32 kBytes. Cada campo é designado por um nome com até 15 caracteres. Nota: maiúsculas e minúsculas são consideradas como diferentes, de modo que CAMPO1, campo1 e Campo1 designariam 3 campos diferentes.

Ver os arquivos “dbtypes.h” e “dblowlevel.h” para as definições do cabeçalho e de campo.

Os arquivos de índice são arquivos binários compostos por um cabeçalho contendo os nomes dos campos usados para compor a chave e uma série de registros contendo o número do registro e o valor da chave.

A extensão do arquivo de dados depende da arquitetura do sistema operacional; em sistemas de 32 bits a extensão será “.i32”, e nos de 64 bits, “.idx”. O espaço reservado para os nomes de campos é de 160 caracteres; portanto, a lista dos campos que compõem a chave deve ter no máximo esse comprimento (incluindo as vírgulas; 10 nomes de campos serão delimitados por 9 vírgulas, por exemplo; assumindo que cada campo tem um nome com 15 caracteres isso significa um total de 159 caracteres).

Ver o arquivo “findex.h” para as definições relevantes para o arquivo de índice.

Índices são gerados em duas etapas: primeiramente o arquivo de dados é percorrido em toda sua extensão, e registros contendo o número do registro no arquivo de dados e a respectiva chave são acrescentados ao arquivo de índice. Ao final deste processo, os registros do índice são ordenados segundo sua chave, utilizando-se o algoritmo quicksort. Dependendo do tamanho do arquivo esta operação pode demorar consideravelmente, e consumir bastante memória, por ser uma função recursiva.

A busca no arquivo de índice é feita com o algoritmo de busca binária, que é bastante rápido, mas não localiza necessariamente o primeiro registro de um bloco de chaves iguais. Para garantir essa condição (essencial à operação de vários módulos do programa), a função de busca retrocede no arquivo de índice até localizar a primeira chave. Na maior parte das situações essa diferença de tempo é imperceptível, mas se o arquivo de dados contém um número muito elevado de chaves iguais (dezenas de milhões ou mais), isso pode impactar negativamente a performance de certas operações.

COMPILAÇÃO:

Os fontes do programa estão no arquivo openreclink-N.N-src.zip (sendo N.N a versão desejada), que deve ser descompactado. Um arquivo make (*makefile*) está incluído, mas para utilizá-lo é necessário instalar as bibliotecas adicionais descritas abaixo. O arquivo deve ser editado para ajustar os diretórios utilizados e selecionar a versão para compilação.

Windows:

Deve-se compilar as bibliotecas Xbase (fontes em <http://linux.techass.com/projects/xdb/>) e wxWidgets (fontes em <http://www.wxwidgets.org>), compilar-se todos os fontes sob os diretórios

“gui”, “nongui” e “wxthings” e linká-los com as bibliotecas.

Para utilizar o makefile é necessário usar as ferramentas GNU: MingW (toolchain) e MSys (simula o ambiente Unix), mais o utilitário wx-config.exe (<http://sites.google.com/site/wxconfig/>). Ver a documentação deste último sobre as variáveis de sistema que devem ser criadas para utilizá-lo.

Linux:

Para compilação sob Linux, são necessários pacotes adicionais:

wxWidgets:

- wx2.8-headers
- libwxgtk2.8-dev
- libwxgtk2.8-dbg
- libwxbase2.8-dev
- libwxbase2.8-dbg

xbase:

- libxbase2.0-dev

ANEXO: GPL Versão 3

GNU GENERAL PUBLIC LICENSE

Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <<http://fsf.org/>>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program--to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the GNU GPL protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the GPL clearly explains that there is no warranty for this free software. For both users' and authors' sake, the GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim

of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the GPL to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the GPL, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the GPL assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS

0. Definitions.

"This License" refers to version 3 of the GNU General Public License.

"Copyright" also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

"The Program" refers to any copyrightable work licensed under this License. Each licensee is addressed as "you". "Licensees" and "recipients" may be individuals or organizations.

To "modify" a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a "modified version" of the earlier work or a work "based on" the earlier work.

A "covered work" means either the unmodified Program or a work based on the Program.

To "propagate" a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To "convey" a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays "Appropriate Legal Notices" to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

1. Source Code.

The "source code" for a work means the preferred form of the work for making modifications to it. "Object code" means any non-source form of a work.

A "Standard Interface" means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The "System Libraries" of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and (b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A "Major Component", in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The "Corresponding Source" for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work's System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding Source.

The Corresponding Source for a work in source code form is that same work.

2. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works

for you must do so exclusively on your behalf, under your direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

3. Protecting Users' Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

4. Conveying Verbatim Copies.

You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

5. Conveying Modified Source Versions.

You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

- a) The work must carry prominent notices stating that you modified it, and giving a relevant date.
- b) The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to "keep intact all notices".
- c) You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of

the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.

d) If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so.

A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an "aggregate" if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's users beyond what the individual works permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

6. Conveying Non-Source Forms.

You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

a) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.

b) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.

c) Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.

d) Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you

or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.

e) Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d.

A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

A "User Product" is either (1) a "consumer product", which means any tangible personal property which is normally used for personal, family, or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, "normally used" refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

"Installation Information" for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in ROM).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or installed. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

7. Additional Terms.

"Additional permissions" are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

- a) Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or
- b) Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or
- c) Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or
- d) Limiting the use for publicity purposes of names of licensors or authors of the material; or
- e) Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or
- f) Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors.

All other non-permissive additional terms are considered "further restrictions" within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the

relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

8. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License (including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

9. Acceptance Not Required for Having Copies.

You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

10. Automatic Licensing of Downstream Recipients.

Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

An "entity transaction" is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who

receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

11. Patents.

A "contributor" is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor's "contributor version".

A contributor's "essential patent claims" are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, "control" includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a "patent license" is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To "grant" such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party. If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2) arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. "Knowingly relying" means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient's use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is "discriminatory" if it does not include within the scope of its coverage,

prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business of distributing software, under which you make payment to the third party based on the extent of your activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

12. No Surrender of Others' Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

13. Use with the GNU Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the GNU Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the GNU Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

14. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the GNU General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the GNU General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the GNU General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the GNU General

Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

15. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

17. Interpretation of Sections 15 and 16.

If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

END OF TERMS AND CONDITIONS