Artigo Original *Infocomunicologia*

Processamento de Dados Tabulares utilizando Ferramentas com Interface em Modo Texto

Tabular Data Processing utilizing Tools with a Text Mode Interface

Procesamiento de Datos Tabulares utilizando Herramientas con Interfaz al Modo de Texto

Gustavo Stein*

* Servidor Público Federal. Bacharel em Informática com Ênfase em Análise de Sistemas. Voluntário da Associação do Centro de Altos Estudos da Conscienciologia (CEAEC).

stein.gustavo@gmail.com

Palavras-chave
Comandos
CSV
GNU/Linux
LibreOffice
Tabelas
Terminal

Keywords
Commands
CSV
GNU/Linux
LibreOffice
Tables
Terminal

Palabras-clave
Comandos
CSV
GNU/Linux
LibreOffice
Tablas
Terminal

Artigo recebido em: 10.03.2017
Aprovado para publicação em: 31.03.2017

Resumo:

O artigo expõe uma possibilidade técnica ao pesquisador que precisar fazer tratamento de grandes quantidades de dados arranjados de modo detalhado em tabelas. O objetivo é compartilhar o método utilizado na geração e na organização das tabelas apresentadas na presente edição da revista *Conscientia*, visando demonstrar uma maneira assertiva de trabalho e, também, auxiliar na detecção e correção de erros nos dados. As tabelas trabalhadas disponibilizam informações gerais dos títulos, autores, especialidades e datas das 850 publicações na forma de artigos, relatos, resenhas, resumos e outros, feitas por 477 autores em 2 décadas de existência da revista *Conscientia*, de 1997 a 2016, contemplando 123 especialidades conscienciológicas.

Abstract:

This article exposes a technical task related to a researcher who needs to treat large amounts of data arranged, in a detailed manner, in tables. The objective is to share the method used in the generation and organization of tables presented in this issue of the journal *Conscientia*, aiming to demonstrate an assertive way of working, and also, to help in the detection and correction of errors in the data. The tables worked on included general information on the titles, authors, specialties and data of the 850 publications, in the form of articles, reports, reviews, abstracts and others, by 477 authors accumulated over the 2 decades the journal *Conscientia* has existed, from 1997 to 2016. These works cover 123 conscientiological specialities.

Resumen:

El artículo expone una posibilidad técnica para el investigador que necesita aplicar gran cantidad de datos colocados, de modo detallado, en tablas. El objetivo es compartir el método utilizado en la generación y organización de las tablas presentadas en la actual edición de la revista *Conscientia*, buscando demostrar una manera afirmativa de trabajo y, también, auxiliar en la detección y corrección de errores en los datos. Las tablas trabajadas ofrecen informaciones generales de: títulos, autores, especialidades y fechas de las 850 publicaciones en forma de artículos, relatos, reseñas, resúmenes y otros, realizadas por 477 autores durante las 2 décadas de existencia de la revista *Conscientia*, de 1997 a 2016, contemplando 123 especialidades concienciológicas.

INTRODUÇÃO

No Léxico de Ortopensatas, Vieira (2014, p. 1.601), expressa:

"* Use a **Tecnologia** a seu favor: faça do computador a sua caixa de pandora e do banco de dados a sua cornucópia".

Alinhado a tal ideia, o autor desenvolveu o trabalho de manuseio eletrônico dos dados e a elaboração das tabelas demonstrativas das publicações realizadas nos 20 anos de existência da revista *Conscientia*.

O objetivo foi possibilitar aos leitores diversas possibilidades de consultas para visão geral e acesso ao conjunto das publicações na revista *Conscientia,* de 1997 a 2016, fruto de trabalhos de pesquisadores e pesquisadoras da Conscienciologia.

O texto visa compartilhar o processo utilizado para o tratamento dos dados e a produção das tabelas dos apêndices desta edição da revista *Conscientia,* demostrando possibilidade técnica ao pesquisador interessado em processar grandes quantidades de dados de maneira rápida, com baixa probabilidade de erros, auxiliando a encontrar e corrigir erros nos dados.

Dessa forma, o autor disponibiliza didaticamente a descrição do que foi feito, de modo a colaborar para que os leitores interessados possam apreender o processo a partir de exemplo prático descrito, contribuindo para o enriquecimento e facilitação na apuração das pesquisas pessoais, a partir de dados levantados.

Devido às particularidades necessárias para obter o resultado, a metodologia utilizada está descrita ao longo das explanações descritas adiante.

A apresentação está feita em 5 seções, expostas em sequência lógica, as quais, em linhas gerais passa por normalização dos dados de uma planilha de dados original, conversão da planilha para modo texto puro e manipulação dos dados com as ferramentas descritas no artigo.

Para melhor entendimento do processo descrito, é útil a visualização das tabelas apresentadas nas páginas anteriores, quando mencionadas, além do *layout* da planilha BDRevistaConscientia.ods descrita ao final da seção IV.

I. Experimentos Iniciais

O autor recebeu a tabela original, encaminhada com o artigo de Flávio Buononato, repassada pela editora da Revista, com a relação das publicações na Revista Conscientia ao longo dos seus 20 anos. A tabela utilizada, denominada *Autores 20 anos,* possuía 1.096 linhas de dados (excluída a linha de título) e 8 colunas assim denominadas, listadas na ordem apresentada na planilha:

1. Autor(a) – o nome completo do autor.
2. Gênero – masculino ou feminino.
3. Tipo de Publicação – artigo, resumo, resenha e outros.
4. Produção – se individual ou em coautoria.
5. Data Publicação – o ano da publicação.
6. Título – o título da publicação.
7. Especialidade – a especialidade da Conscienciologia na qual se insere a publicação.
8. Dados Complementares – O volume, número e evento associado à publicação, se for o caso, na forma Vol. 99 No. 9 - Nome do Evento.

A ideia inicial era separar a tabela geral em duas outras: uma agrupada e ordenada por especialidades contendo ainda as colunas título, dados complementares renomeada para edição - evento e tipo de publicação e também outra tabela similar às especialidades, porém substituindo a coluna especialidade por autor(a) agrupada e ordenada por estes.

Após avaliação, em conjunto com o autor Flávio Buononato, decidiu-se colocar no Apêndice 1 a tabela original classificada alfabeticamente pelo nome do autor e em ordem cronológica de publicação. Na tabela publicada foram suprimidos os campos Gênero, Data de Publicação e Dados Complementares. Este último foi dividido e suprimida a informação do evento associado ficando denominado Edição. Estas supressões foram realizadas para diminuir a quantidade de páginas necessárias e os dados restantes considerados suficientes sem prejudicar o resultado final.

A tabela do Apêndice 1 possui correções realizadas durante a etapa de processamento de dados para produção das demais tabelas do Apêndice 2. Basicamente foram corrigidos erros de grafia de especialidades da Conscienciologia e normalizada a acentuação de nomes dos autores.

As tabelas do Apêndice 2 foram produzidas com o objetivo de sintetizar os dados da tabela original para facilitar a busca pelo pesquisador.

A natureza da tabela do Apêndice 1, agrupando as diversas informações listadas, ocasiona repetições inevitáveis, a exemplo de publicações realizadas em coautoria. Nestas, ocorrem a repetição de tantas linhas do título da mesma publicação quantos autores participaram da respectiva publicação. Desta forma, um artigo realizado com a partição de 3 autores aparecerá listado 3 vezes na tabela.

Uma vez que não estava totalmente claro no primeiro momento como seriam produzidas as tabelas do Apêndice 2, realizaram-se vários experimentos com a planilha das publicações. E, houve bastante retrabalho. Listar todas estas experiências neste artigo o deixaria muito extenso e poderia confundir o leitor. Na sequência apresenta-se o processo de produção do resultado final e onde for apropriado serão tecidos comentários pertinentes. Serão ainda apresentadas recomendações para evitação de retrabalhos.

II. Preparação ao Processamento dos Dados

Utilizando o *software* de planilhas LibreOffice Calc, foi criada a partir do arquivo da planilha original, que possuía várias planilhas e gráficos incorporados, um arquivo com cópia simples da planilha denominada *Autores 20 anos*. Esta planilha foi salva com o nome BDRevistaConscientia.ods no formato .ODS nativo do LibreOffice Calc.

Na primeira ocasião foram copiados somente os dados desejados, contudo a experiência demonstrou a importância de copiar todos os dados originais, pois posteriormente alguma necessidade pode fazer com que alguns dados não copiados possam ser necessários.

Também é importante não se trabalhar sobre os dados originais pois qualquer alteração inadvertida poderá não ser de possível reversão posteriormente. Foi acrescentado na tabela copiada um campo (coluna) denominado Seq. (abreviatura de Sequência) contendo a sequência original dos registros na tabela original de 1 a 1.096.

A intenção deste campo é poder retornar os registros à ordenação original a qualquer momento necessário para analisar os dados na forma original e também, quando necessária a criação de subtabelas, possuir um campo que permita garantir que a colagem da subtabela posteriormente na tabela agrupadora preserve o mesmo ordenamento dos dados.

Fazer os processamentos de dados necessários para produção das tabelas do Apêndice 2 utilizando-se somente o *software* de planilhas, seria muito trabalhoso e propenso a erros. Desta forma, aproveitando a experiência anterior do autor com os comandos utilizados no sistema operacional UNIX e sua variante bem conhecida, o GNU/Linux, optou-se por utilizar os comandos disponíveis neste, para processamento de textos.

Este processo é descrito a partir da seção IV. Vários comandos foram executados originalmente de modo separado, inclusive para testes, contudo foram reescritos de forma integrada visando reduzir a quantidade final de comandos necessários.

III. Breve Descrição do Sistema Operacional GNU/Linux

Sistema Operacional é o *software* básico instalado em qualquer computador, sem o qual o usuário não consegue utilizá-lo. Exemplos de sistemas operacionais: Android, iOS, Mac OS X, Windows 10.

O sistema operacional GNU/Linux é conhecido geralmente somente pelo nome Linux, ou pelo nome de uma de suas distribuições a exemplo de Debian, Fedora, Mint, Redhat, Ubuntu e outras.

Contudo o sistema é derivado de 2 projetos, um denominado GNU, criado em 1984, por Richard Stallman (1953-), anterior ao Linux, criado em 1991 por Linus Torvalds (1969-).

O Linux é somente o núcleo do sistema operacional, a parte invisível responsável pela comunicação entre o *hardware* do computador e o *software* visível utilizado pelo usuário, além do gerenciamento dos recursos do *hardware* e *software* utilizados.

A parte visível do sistema operacional é em boa parte *software* GNU ou derivado deste. Por isto, para mostrar as origens do projeto, o nome mais adequado do sistema operacional é GNU/Linux.

Ao referir-se ao sistema operacional denominando-o somente Linux, o usuário está simplesmente ignorando a parte visível do sistema operacional, com a qual ele interage diretamente, por mais estranho que pareça.

O GNU/Linux possui aplicativos de modo gráfico a exemplo do Google Chrome, LibreOffice, Mozilla Firefox e tantos outros, e também possui comandos executados em modo terminal conhecido também por *prompt* ou linha de comandos (similar àquela tela preta geralmente com caracteres brancos exibida ao se executar o programa cmd no Windows).

O antigo MS-DOS e o modo terminal do Windows mesmo nas suas versões mais recentes foram desenvolvidos a partir de um subconjunto de comandos criados originalmente no UNIX nos idos da década de 1970.

O GNU/Linux foi desenvolvido inspirado no UNIX que possuía, originalmente, somente modo terminal. Embora pareça ser ultrapassado, o modo terminal possui comandos, alguns bastante simples e outros extremamente poderosos que combinados nos permitem fazer processamentos muito mais rápidos que os aplicativos gráficos.

Na realidade até hoje (Ano-base: 2017), não se conseguiu desenvolver aplicativos gráficos que possuam conjuntamente a mesma flexibilidade e poder de processamento dos comandos do terminal do UNIX.

IV. Preparação e Conversão dos Dados da Planilha para modo Texto Puro

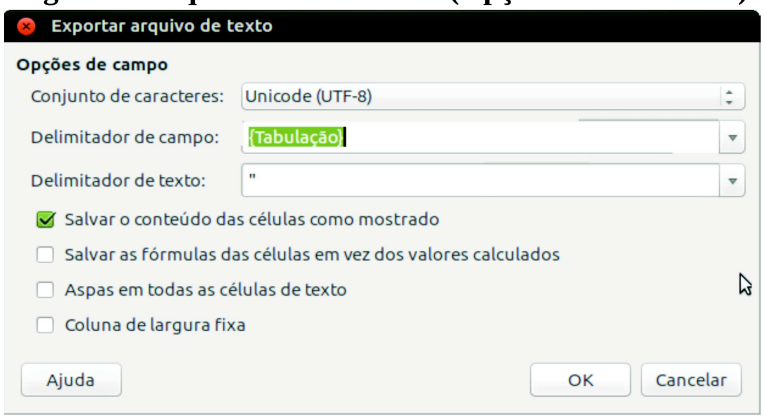
Os comandos de terminal do GNU/Linux nativamente não entendem o formato de dados utilizado pelos aplicativos de planilha eletrônica a exemplo do LibreOffice Calc ou Microsoft Excel.

Estas planilhas são armazenadas em formatos binários compactados ininteligíveis aos variados comandos. Para isto é preciso primeiro realizar-se um processo de conversão de dados.

O processo de conversão de dados foi realizado a partir do *software* LibreOffice Calc utilizando o comando Arquivo → **Salvar como…**, selecionando-se o formato .CSV (*comma separated values*) para saída,

porém utilizando-se o caractere de tabulação como separador das colunas ou campos da planilha. A seguir a Figura 1 mostra a janela exibida após a seleção do formato .CSV.

Figura 1: Arquivo - Salvar como (Opções do filtro CSV)



O formato de arquivo .CSV é um tipo de arquivo eletrônico de computador em que somente é utilizado texto em formato bruto a exemplo do .TXT empregado pelo aplicativo Notepad no Windows. No formato .CSV, a primeira linha contém o título dos campos armazenados no arquivo, separados por um caractere separador. As demais linhas contêm os dados listados linha a linha separados pelo mesmo separador empregado na primeira. Os campos nas linhas de dados estão armazenados na mesma ordem em conformidade à primeira. Todas as linhas do arquivo são terminadas pelo caractere de final de linha, que varia conforme o sistema operacional utilizado pelo usuário, seja BSD, GNU/Linux, Mac OS, Windows ou outros.

Na especificação original, o próprio nome do formato sugere, é utilizado o caractere vírgula (*comma* em inglês) para separação de campos, contudo em sua especificação admite-se utilizar outros caracteres. Conforme experiências do autor com manipulações de outros arquivos armazenados neste formato, o caractere mais confiável a ser empregado para evitar-se erros na manipulação dos dados é o caractere de tabulação, o mesmo gerado pela tecla TAB do teclado do computador. Segue abaixo exemplos de arquivo .CSV, o primeiro com tabulação e o segundo com vírgulas. O texto é mostrado em fonte mono espaçada pois é a aparência como o texto é visualizado no modo terminal. Os comandos mostrados mais a frente utilizarão também esta fonte para ficar evidenciado serem comandos visualizados em modo terminal.

```
Título      Edição
Acoplamentarium: Experimentologia Grupal Avançada     Vol. 8 No. 2
Atualização do Cadastro – Relato de um Trabalho Multidimensional  Vol. 10 No. 1
Empresas Conscienciológicas: Hipóteses de Trabalho    Vol. 10 No. 1
```

```
Título,Edição
Acoplamentarium: Experimentologia Grupal Avançada,Vol. 8 No. 2
Atualização do Cadastro – Relato de um Trabalho Multidimensional,Vol. 10 No. 1
Empresas Conscienciológicas: Hipóteses de Trabalho,Vol. 10 No. 1
```

Como se observa, diferente de uma planilha, os campos ficam emendados um no outro somente separados pelo separador definido. Diferentes caracteres, a exemplo da vírgula, podem compor parte dos dados de um campo de um arquivo .CSV. Em alguns casos, ao fazer um processo de conversão, similar ao descrito,

é aconselhável antes verificar se o caractere separador não está sendo utilizado nos dados armazenados no arquivo a ser convertido, pois se o for, os procedimentos descritos mais adiante irão falhar ou ocasionar resultados imprevisíveis.

Para permitir o adequado processamento dos arquivos de modo texto com os comandos do GNU/Linux é adequado que os dados estejam o mais normalizados possíveis. Exemplos de problemas:

1. Linhas divididas na mesma célula com o caractere ENTER. Todo o processamento dos dados dos comandos de terminal são realizados por linha, utilizando-se como separador de linha o caractere ENTER representado por diferentes caracteres conforme o sistema operacional instalado no computador do usuário. As células que possuem texto divido em múltiplas linhas geram problemas de processamento. Felizmente os próprios comandos do GNU/Linux nos mostram quando existem problemas, pois as saídas esperadas ficam divididas, incompletas ou apresentando problemas de visualização. Para eliminar o problema das linhas divididas o melhor a fazer, no caso do LibreOffice Calc, é a eliminação dos caracteres de final de linha, utilizando-se o comando Editar → **Localizar e substituir...** ou Ctrl+H, utilizando-se as seguintes opções:

Em Localizar: digitar somente os dois caracteres \n (barra invertida e n).

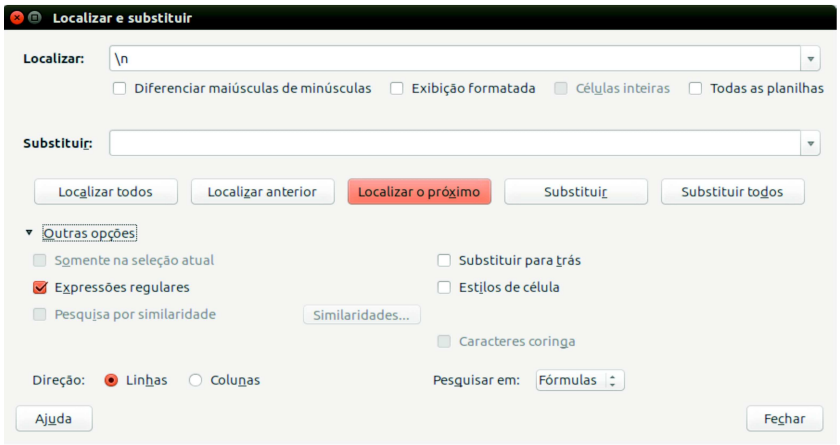
Em Substituir: deixar em branco.

Clicar em **Outras opções** e marcar a opção **Expressões regulares**.

Clicar no botão **Substituir todos** e depois no botão Fechar duas vezes.

A figura 2 ilustra a janela para a operação de eliminação dos caracteres de fim de linha.

Figura 2: Remoção dos finais de linha das linhas divididas



Se o editor de planilhas utilizado não possuir estas opções, uma forma alternativa é alargar bastante as colunas da planilha de forma a forçar as linhas divididas a aparecerem pois estas permanecerão divididas não importa o quanto se aumente a largura da coluna. É bom lembrar que esta segunda opção é mais demorada e sujeita a erros.

2. Espaços em branco excedentes. Outra normalização utilizada foi a eliminação de espaços em branco excedentes desnecessários. Foram substituídos dois espaços em branco por um, duas vezes, de forma a que em toda a planilha existissem somente um espaço em branco separando duas palavras. Isto se faz necessário para facilitar os procedimentos de classificação, ordenação, agrupamento e contagem de itens iguais, que com os espaços excedentes seriam considerados diferentes. No caso presente somente duas substituições foram suficientes para eliminação de todos os caracteres em branco desnecessários.

Uma vez normalizados os dados, a planilha foi salva e então convertida no formato .CSV, conforme mostrado, com o comando *Salvar Como...* com o nome BDRevistaConscientia.csv para os procedimentos posteriores. Na forma final, a planilha *BDRevistaConscientia.ods* ficou com os seguintes campos na ordem apresentada: *01/A. Autor(a), 02/B. Tipo de Publicação, 03/C. Produção, 04/D. Título, 05/E. Especialidade, 06/F. Edição, 07/G. Código, 08/H. Núm., 09/I. Evento, 10/J. Dados Complementares, 11/K. Seq., 12/L. Gênero e 13/M. Data Publicação*. As letras indicadas após os números são as referências às colunas na planilha. Os comandos descritos posteriormente utilizarão a numeração de campos para referência para fins didáticos embora possam não ter sido os casos reais utilizados. Estão enumerados campos ainda não descritos contudo serão mencionados quando necessários.

V. Manipulação dos Dados da Planilha em modo Texto Puro

Para se fazer o processamento do arquivo .CSV convertido a partir da planilha original foram utilizados os comandos descritos abaixo, todos executados a partir do modo terminal do GNU/Linux, pois todos os comandos descritos são diretamente executados deste modo não possuindo interfaces gráficas, somente saídas de textos. Os comandos foram executados no modo terminal posicionado na pasta onde os arquivos da revista estavam sendo criados. Os comandos são simples e podem ser combinados para processamentos mais elaborados. Todos estes comandos só funcionam em arquivos que estiverem em formato texto, conforme já exposto. No momento oportuno serão explicados os detalhes necessários. Segue a relação de comandos empregados com uma breve descrição da sua funcionalidade geral:

cut – corta o conteúdo de um arquivo extraindo só as colunas, campos ou quantidade de caracteres especificada.

echo – exibe ou imprime na tela o texto ou resultado de um processamento anterior.

for – comando de repetição. Utilizado para se repetir comandos quando necessário.

grep – filtra o conteúdo de um arquivo conforme os critérios fornecidos.

less – paginador de arquivos. Permite exibir o conteúdo de um arquivo página a página. O padrão dos comandos do GNU/Linux é exibir todo o conteúdo de um arquivo resultado de processamento na tela, o que pode demorar se houver muita informação a ser exibida. Utilizado para evitar a rolagem desnecessária de texto na tela ao se examinar o conteúdo de um arquivo.

sed – editor de textos de fluxo de dados. Permite realizar substituições e transformações em textos, entre elas a troca de uma sequência de texto por outra de forma automatizada. É a versão programática do comando *Localizar e substituir* existente nos editores de planilha e textos, porém muito mais poderoso.

sort – classifica as linhas de um arquivo. Por padrão em ordem alfanumérica.

tr – troca o(s) caractere(s) especificado(s) contidos em um arquivo por outro(s) caractere(s) conforme especificado. Diferentemente do sed, o tr troca somente um caractere individual por outro. Quando necessário trocar um caractere por dois ou mais, ou vice-versa, o sed é obrigatório.

uniq – elimina as linhas repetidas contíguas de um arquivo. Para que isto seja possível é necessário primeiro classificar as linhas com o comando sort.

O modo terminal nos sistemas operacionais é executado em um interpretador de comandos. Este interpretador é mais um comando disponível dentre os vários existentes. No GNU/Linux e MAC OS X, o interpretador de comandos padrão é o bash, e no Windows é o CMD.EXE. Cada interpretador de comandos pos-

sui associado uma linguagem de programação. No caso do bash a linguagem utilizada é chamada de Shell. Serão utilizados alguns comandos básicos de linguagem de programação e serão explicados quando utilizados. Todos os comandos neste artigo foram executados na distribuição GNU/Linux denominada Ubuntu na versão 16.04. Os comandos podem funcionar de forma igual ou similar no MAC OS X, e, no Windows 10 versão *Anniversary Update* ou superior, se instalado o pacote *Windows Subsystem for Linux*.

Para extrair a relação completa das especialidades existentes no arquivo BDRevistaConscientia.csv utiliza-se o comando abaixo:

```
cut -f5 BDRevistaConscientia.csv | grep -v Especialidade | sort | uniq > EspecialidadesOrdenadas.txt
```

Embora o comando anterior apareça dividido em duas linhas, deve ser digitado em somente uma linha no terminal utilizando-se somente espaços para separar uma expressão de outra.

O argumento **-f5** instrui o **cut** para cortar somente o campo (*field*) indicado (no caso 5) do arquivo BDRevistaConscientia.csv. Por padrão qualquer saída de um comando é jogada na tela. Neste caso seria listado na tela o título Especialidade e nas demais linhas as especialidades de todos os tipos de publicação de todas as revistas publicadas. Isto não ocorre devido a presença do caractere (|) conhecido como pipe.

O caractere pipe (|) é um caractere especial utilizado na linha de comando para criar um canal de comunicação entre a saída do comando anterior (cut) que iria para a tela e a direciona ao comando posterior (grep) que por padrão esperaria a sua entrada a partir do teclado.

O comando **grep** é utilizado como filtro. Neste caso o argumento **-v** fornecido ao **grep** solicita a eliminação do argumento posterior, dos dados recebidos. O argumento fornecido após o **-v** é eliminado tantas vezes quantas existentes no arquivo (neste caso somente uma vez pois se trata do título da coluna).

A saída do **grep** (todas as demais especialidades do arquivo original) é então direcionada ao comando **sort** para ordenar todas as especialidades em ordem alfabética. A saída do comando **sort** é passada ao comando **uniq** para eliminar todas as repetições localizadas em linhas contíguas deixando somente uma ocorrência de cada especialidade.

A saída do comando **uniq** por padrão seria a tela, contudo o caractere maior (>) a direciona para o arquivo denominado EspecialidadesOrdenadas.txt. O arquivo é criado com a lista de todas as especialidades localizadas no arquivo BDRevistaConscientia.csv.

Como resultado do comando anterior, após análise do arquivo criado, foram encontradas as seguintes especialidades com erros de grafia ou forma: *Amparalogia, Autoconcienciometrologia, Autoexperiementologia, Autoexperimentologia, Autopesquisologia, Comoeticologia, Epstemologia, Expertimentologia e Paradiplomaticologia*. As devidas correções foram realizadas no arquivo da planilha BDRevistaConscientia.ods, esta novamente convertida para o formato .CSV. Depois das correções restaram, no total, 123 Especialidades da Conscienciologia.

Processo similar foi realizado para os autores a fim de normalizar os nomes. Foram localizadas e trocadas as partículas **De** para **de** em alguns nomes de autores e foram acentuados nomes de autores que estavam sem acentos conforme as regras do português. Não foram alterados nomes de autores conhecidos que não utilizam normalmente a acentuação. Neste sentido este autor sugere, a criação de um banco de dados oficial de autores com a grafia do nome desejada a ser divulgado, por exemplo, no *site* do ICGE.

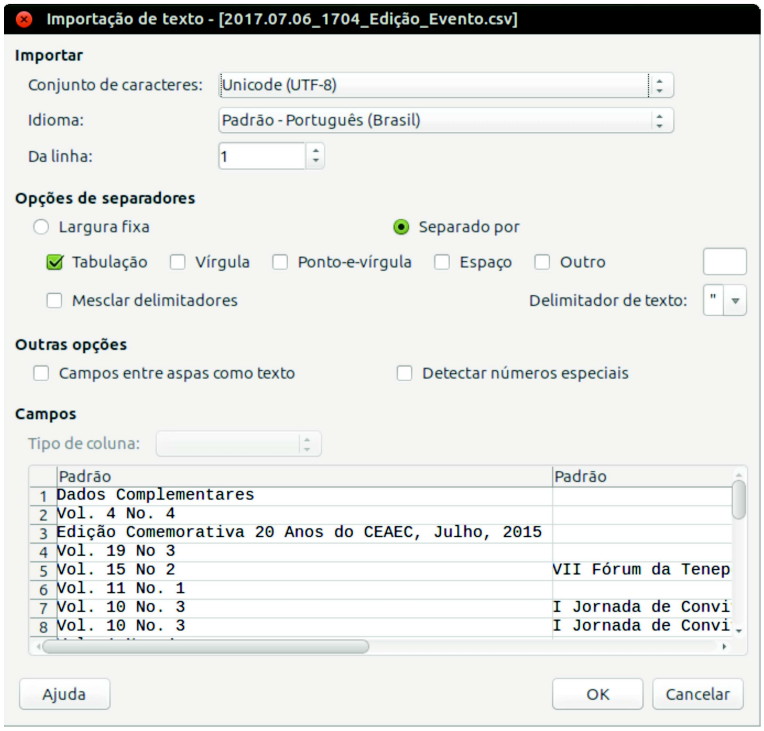
Para separar a coluna Dados Complementares em Edição e Eventos separadamente foi executado o comando seguinte no terminal:

```
cut -f10 BDRevistaConscientia.csv | sed 's/ - /\t/g' > Edição_Evento.csv
```

O argumento **-f10** solicita ao **cut** para selecionar o décimo campo (coluna) do arquivo BDRevistaConscientia.csv, lembrando que as colunas estão separadas pelo caracter tabulação conforme já recomendado. A saída do comando é repassada com o caractere pipe (|) ao comando **sed** com o argumento '**s/ - /\t/g**'. O argumento entre aspas simples repassado ao **sed** é um *script* (roteiro) da ação a ser realizada sobre os dados recebidos. No caso a letra (**s**) significa substituir. A barra (/) é utilizada para separar argumentos do *script*. No caso, o primeiro argumento entre barras é o <ESPAÇO><HÍFEN><ESPAÇO> que será substituído pelo próximo argumento entre as próximas duas barras, no caso **\t**. O **\t** é uma representação para o caractere de tabulação. Após a última barra (/) antes da última aspas simples temos a letra (**g**), que significa global, ou seja, a substituição do texto ' - ' por '**\t**' será realizada em todos os dados recebidos. Se a letra **g** não fosse colocada, por padrão, o **sed** só substituiria a primeira ocorrência encontrada nos dados recebidos.

A partir de uma planilha vazia no LibreOffice Calc, ao abrir o arquivo Edição_Eventos.csv, o Calc irá abrir o assistente Importação de texto. Em **Opções de separadores** devem estar selecionadas as seguintes opções: **Separado por** e Tabulação. Na figura 3, exposta a seguir, é exibido um exemplo da janela do assistente de Importação de texto.

Figura 3: Assistente de Importação de texto



Ao clicar em OK as colunas Edição e Eventos estarão separadas na planilha carregada. Lembrando de trocar o título de *Dados Complementares* que estará situada na coluna mais à esquerda por Edição (A) e Eventos (B) basta selecionar as duas colunas separadas e colá-las no arquivo BDRevistaConscientia.ods nas colunas 06 (F) e 09 (I) já anteriormente reservadas conforme citado no *layout* anteriormente. Para assegurar-se que os dados não fossem alterados de ordem entre a extração do campo 10 (Dados Complementares) mostrada acima e a colagem dos dados na tabela BDRevistaConscientia.ods o campo **Seq.** poderia ter sido extraído juntamente, substituindo-se somente o argumento **-f10** do comando **cut** mostrado acima por **-f10,11**.

Neste caso ao abrir-se o arquivo Edição_Eventos.csv no LibreOffice Calc a planilha exibiria 3 colunas. Uma vez alterado o arquivo BDRevistaConscientia.ods novamente é preciso convertê-lo para o formato .CSV a fim de se possuir o arquivo .CSV com as novas alterações realizadas para os processamentos seguintes.

A partir deste momento o campo Edição da planilha BDRevistaConscientia se apresentava nas formas 'Vol. 9 No. 9' ou 'Vol. 99 No. 9' conforme o número e volume da edição com os volumes de 1 a 9 sem o 0 à esquerda. A fim de facilitar a classificação das edições, melhorar a estética de apresentação e diminuir a largura do campo foi executado o comando abaixo:

```
cut -f6,11 BDRevistaConscientia.csv | sed 's/Vol\./V./g' | sed 's/No\./N./g' | sed
's/No /N. /g' | sed 's/Suplemento/Supl./g' | sed -E 's/V. ([1-9]) /V. 0\1 /g' | sed
's/Edição.Comemorativa.20.Anos.do.CEAEC,.Julho,.2015/EC20AC */g' > Edição.csv
```

Conforme já exposto anteriormente, o comando é digitado em somente uma linha utilizando espaços para separar um comando do outro. Os espaços antes e depois do caractere (|) utilizado para concatenação dos comandos são opcionais, aqui mostrados somente para melhor visualização. Já, entre o comando **sed** e a primeira aspas simples, é obrigatório a utilização de um espaço em branco. Para resumir, o comando anterior realiza o seguinte: Extrai o 6º campo (Edição) e o 11º (Seq.) do arquivo da tabela geral e troca os textos abaixo mostrados na coluna *Texto original* pelo conteúdo correspondente na coluna *Texto substituído* todos mostrados entre aspas simples para melhor visualização.

Texto original	**Texto substituído**
'Vol.'	'V.'
'No.'	'N.'
'No '	'N. '
'Suplemento'	'Supl.'
'V. 1a9'	'V. 01a09'
'Edição Comemorativa 20 Anos do CEAEC, Julho, 2015'	'EC20AC *'

O comando anterior possui 7 comandos encadeados, iniciando-se pelo **cut** e utilizando-se o **sed** 6 vezes seguidas para realizar as substituições acima. Obviamente poderia ter sido usado o comando de busca e troca do editor de planilha de preferência contudo seria necessário realizar-se, no total, 14 substituições para obter-se o mesmo resultado, pois foram trocados os números dos volumes de 1 a 9 para 01 a 09. Também seria preciso cuidado para marcar somente a coluna Edição na planilha, senão poderia ser substituído texto não desejado, por exemplo do campo 10 (Dados complementares).

Em relação às substituições acima merece ser destacado a terceira 'No ' por 'N. '. Esta foi necessária ser realizada pois em algumas edições estava faltando o ponto (.) logo após a letra (o). O padrão das demais edições era possuir ponto (.) logo após a expressão No.

Merece explicação o comando **sed -E 's/V. ([1-9]) /V. 0\1 /g'**. O argumento **-E** fornecido ao **sed** ativa o modo de expressões regulares estendido. Este modo é necessário para que seja entendida a expressão **([1-9])**. Os parênteses desta expressão marcam um ponto que será definido como uma variável, ou seja, o que estiver dentro do parênteses será salvo para uso posterior. É possível termos até nove conjuntos de parênteses a serem atribuídos a variáveis identificadas pelo conjunto \1 a \9. Os conjuntos de parênteses são numerados automaticamente da esquerda para direita. No caso em questão estamos utilizando somente um conjunto. A expressão dentro dos parênteses **[1-9]** instrui o **sed** para esperar naquela posição dígitos de 1 a 9, qualquer

outro caractere que for localizado em uma sequência que se inicie por ‘V. ’ será ignorado e não salvo. No entanto se for um dígito de 1 a 9 ele será salvo e utilizado quando desejado. Verificando na linha acima após a expressão ‘**V. ([1-9])** /’ temos ‘**V. 0\1** /’. Isto instrui o **sed** a trocar ‘V. 1 ’ a ‘V. 9 ’ por ‘V. 01 ’ a ‘V. 09 ’. O que foi encontrado na posição marcada pelos parênteses, no caso, dígitos de 1 a 9, pois assim foi especificado, será colocado no lugar da expressão ‘**\1**’.

O campo Seq. só foi extraído para garantir a sequência quando da colagem dos dados na planilha BDRevistaConscientia.ods. Poderia ter sido ignorado. No comando acima o caractere separador da Edição Comemorativa mostrado foi o ponto (.). Por padrão o comando **sed** e a grande maioria dos comandos do GNU/Linux interpreta o ponto como um padrão para qualquer caractere. Poderia ter sido utilizado o espaço porém o ponto aqui foi utilizado para melhor visualização da quantidade de caracteres. Após o último comando **sed** utiliza-se o caractere maior (>) para direcionar a saída ao arquivo Edição.csv. Uma vez aberto este arquivo no *software* de planilha e sobrepostos os dados na coluna de Edição, o arquivo BDRevistaConscientia.ods foi salvo e então classificado pelo campo Autor(a) e Edição. A partir desta versão da planilha, foram copiadas as colunas A a F (01 a 06 conforme *layout* mostrado anteriormente) e coladas no editor LibreOffice Writer para montagem da tabela geral do Apêndice 1 desta revista.

Para a produção das tabelas do Apêndice 2 utilizou-se algumas transformações nos dados. Em reunião com a editora da revista, decidiu-se mudar o tipo de algumas publicações conforme mostrado a seguir:

Antiga classificação	**Nova classificação**
Autopesquisa da Consciência	Artigo
Correspondências	Cartas
Fundamentos da Conscienciologia	Artigo
Informações	Informativo
Opinião	Artigo
Visão Histórica	Artigo

Esta mudança foi realizada porque os tipos de publicação da coluna *Antiga classificação* não são mais utilizados. As alterações de Correspondências para Cartas e Informações para Informativo foram realizadas também na tabela do Apêndice 1. Uma vez realizadas as alterações com a substituição dos textos anteriores no próprio LibreOffice Calc, os dados da planilha BDRevistaConscientia.ods foram classificados de forma alfanumérica pelos campos *Tipo de publicação* e Título. Foram ainda adicionadas duas novas colunas denominadas Código (coluna G) e Núm. (coluna H), respectivamente os campos 07 e 08 conforme o *layout* mostrado anteriormente. O campo Código foi gerado de maneira automatizada para cada tipo de publicação da forma exposta na tabela a seguir:

Tipo de publicação	**Códigos**
Artigo	A001 a A682
Cartas	C1 a C7
Editorial	E01 a E83
Entrevista	Ent1 e Ent2
Informativo	I1 a I3
Pesquisa Laboratorial	P01 a P19
Relato	Rlt01 a Rlt34
Resenha	Rsh1 a Rsh9
Resumo	Rsm01 a Rsm19

Na primeira linha de cada tipo de publicação na coluna Núm. (H), foi colocado o número 1 e nas demais linhas subsequentes na mesma coluna foi colocada a fórmula =SE(D3=D2;SE(F3=F2;H2;H2+1);H2+1) de forma que o número da coluna H fica inalterado somente quando ambos Título (coluna D) e Edição (coluna F) são iguais aos da linha anterior.

Em qualquer outra situação o número é incrementado de 1. Isto foi necessário devido à repetição dos títulos dos artigos realizados em coautoria. Na primeira linha de dados na coluna Código por sua vez foi colocada a fórmula =(CONCATENAR("A";TEXTO(H2;"000"))) de modo a concatenar o número da coluna H com o texto **A** na forma A999, onde 999 é o número sequencial do artigo em questão. Fórmulas similares foram criadas para todos os tipos de publicação.

Por ser necessário fazer novas operações de manipulações de dados sobre a planilha BDRevistaConscientia.ods esta foi salva com novo nome na forma AAAA.MM.DD_BDRevistaConscientia.ods onde AAAA é o ano com 4 dígitos, MM é o mês com 2 dígitos e DD é o dia com 2 dígitos para preservar as fórmulas se necessário realizar novas operações. O arquivo original BDRevistaConscientia.ods foi salvo novamente no formato .CSV de modo a propositadamente transformar as fórmulas das colunas Código e Núm. em códigos textuais / numéricos respectivamente. Abrindo-se novamente a planilha .CSV no LibreOffice Calc ela foi salva novamente no formato .ODS com o nome original BDRevistaConscientia.ods desta forma eliminando as fórmulas anteriormente criadas.

A tabela anterior relacionando os tipos de publicação e códigos foi copiada ao Apêndice 2 desta revista e denominada Tabela 1.

Copiando-se as colunas Título, Edição e Código da planilha BDRevistaConscientia.ods e colando-se no LibreOffice Writer foi criada a Tabela 2 do Apêndice 2 que relaciona em ordem alfanumérica os Códigos e Títulos das publicações.

Com a sequência de comandos a seguir será gerada a tabela de Especialidades e Publicações denominada EspecialidadesAgrupadas.csv, a ser utilizada para geração da Tabela 3 do Apêndice 2. São mostrados primeiramente os comandos e em seguida a explicação dos mesmos.

```
cut -f5,7 BDRevistaConscientia.csv | grep -v Especialidade | sort | uniq > Especialidades.csv
```

O comando acima extrai os campos 5 (Especialidade) e 7 (Código) do arquivo geral de dados, exclui a linha do título, classifica as linhas extraídas, elimina as repetidas e salva o resultado no arquivo Especialidades.csv.

```
cut -f1 Especialidades.csv | sort | uniq > EspecialidadesOrdenadas
```

O comando acima extrai o primeiro campo (Especialidade) do arquivo Especialidades.csv, classifica as especialidades, elimina as repetições e salva a lista das especialidades individuais no arquivo EspecialidadesOrdenadas.

```
echo -e "Especialidade\tPublicações" >  EspecialidadesAgrupadas.csv
```

O comando acima cria o arquivo EspecialidadesAgrupadas.csv com o título Especialidade e Publicações separadas por uma tabulação, o separador recomendado anteriormente. O argumento **-e** fornecido ao comando **echo** é utilizado para ativar o reconhecimento pelo **echo** das expressões especiais iniciadas pela contrabarra a exemplo da **\t** que é interpretada como tabulação. Sem este argumento seria gravado no arquivo os caracteres contrabarra e **t** em vez da tabulação.

```
for i in $(cat EspecialidadesOrdenadas); do echo -n $i >> EspecialidadesAgrupadas.csv;
echo -ne "\t" >> EspecialidadesAgrupadas.csv; publicacoes=$(grep $i Especialidades.csv|
cut -f2|sort) ; echo $publicacoes | sed 's/ /, /g'  >> EspecialidadesAgrupadas.csv; done
```

O comando anterior é um dos mais complexos utilizados, por envolver uma estrutura de repetição, o comando **for**. Serão explicadas suas partes componentes. Tudo que estiver entre o comando **do** e **done** seguintes ao comando **for** será repetido tantas vezes quantos argumentos forem repassados ao comando **for**.

O comando **cat** (abreviação para *concatenate*) lê todo o conteúdo do arquivo EspecialidadesOrdenadas e em condições normais exibiria o conteúdo na tela. No entanto, por ser executado dentro de parênteses precedido pelo caractere cifrão ($) que precisa estar diretamente ao lado do abre parênteses sem espaços, faz com que o conteúdo do arquivo seja repassado ao comando **for** que espera uma lista de argumentos a serem repetidos, no caso 123 especialidades.

O caractere **i** seguinte ao comando **for** é utilizado na função de variável para armazenar uma especialidade por vez a cada iteração do comando **for**. Na primeira iteração a variável **i** conterá a especialidade Acoplamentologia e na última Voluntariologia, respectivamente a primeira e última especialidades existentes no arquivo EspecialidadesOrdenadas.

O comando **echo -n $i** exibe, por padrão na tela, a especialidade contida na variável **i** e o argumento **-n** fornecido ao comando **echo** solicita a este que não coloque o caractere de final de linha ao imprimir a variável. O padrão do comando **echo** é sempre colocar um caractere ENTER ao final de cada argumento a ser exibido. O argumento **-n** muda este comportamento.

Por existirem os caracteres >> após o comando **echo** em vez do resultado ser exibido na tela, será acrescentado no arquivo EspecialidadesAgrupadas.csv. Nos comandos anteriores foi utilizado somente um caracter maior (>). No entanto, neste caso, são empregados 2 caracteres maior (>) lado a lado sem espaços entre si. Os dois caracteres juntos são interpretados como anexação ao arquivo já existente. Se fosse colocado somente um caractere o conteúdo anterior existente seria eliminado do arquivo e perderíamos todo o conteúdo anterior salvo.

O comando **echo -ne "\t"** é utilizado acima para gravar um caractere de tabulação dentro do arquivo EspecialidadesAgrupadas.csv. Conforme mostrado na Tabela 3 do Apêndice 2, em cada linha da tabela temos na primeira coluna a especialidade e na segunda a lista de todos os códigos das publicações referentes àquela especialidade separados por vírgula. Cada comando individual integrado ao comando maior anterior está montando parte deste arquivo.

O comando **publicacoes=$(grep $i Especialidades.csv|cut -f2|sort)** extrai do arquivo Especialidades.csv todas as linhas que contenham a especialidade armazenada na variável **i** através do comando **grep** que utiliza o argumento **$i** (a especialidade) como filtro e exibe somente as linhas da especialidade em questão, passando-as ao comando **cut** que extrai o segundo campo (os códigos das publicações daquela especialidade), passa o resultado ao comando **sort,** que as classifica, e o resultado é então armazenado na variável **publicacoes**. Neste momento a linguagem Shell nos faz uma gentileza muito bem-vinda. São trocados automaticamente os caracteres de final de linha (ENTER) por espaços em branco.

Antes do leitor pensar que a grafia da variável **publicacoes** acima está incorreta, fica aqui um alerta: na grande maioria das linguagens de programação as variáveis **não** podem ser acentuadas.

O comando **echo $publicacoes** é utilizado para “exibir” o conteúdo armazenado na variável **publicacoes** a ser transferido com o caractere pipe (|) ao comando **sed** que com o argumento **'s/ /, /g'** realiza a troca dos espaços em branco separadores dos códigos por vírgula e espaço. Os caracteres >> acrescentam o resultado ao arquivo EspecialidadesAgrupadas.csv.

Ao chegar ao comando **done** o comando **for** entende que a estrutura de repetição chegou ao seu fim e então o comando **for** volta ao comando seguinte ao **do** utilizando o próximo valor de especialidade disponível até todas as especialidades serem processadas.

A geração da Tabela 4 do Apêndice 2 (autores publicados na revista *Conscientia*) é um pouco mais complexa que o processo para produção da Tabela 3 (de Especialidades) porque os nomes dos Autores contêm espaços em branco entre cada parte do nome. Especialmente o comando **for** falharia se não fosse tomado o cuidado descrito a seguir. A sequência de comandos é muito similar à produção da tabela de Especialidades; serão realizados comentários somente em relação às diferenças do processo anterior. De forma similar serão apresentados os comandos e sua explanação.

```
cut -f1,7 BDRevistaConscientia.csv | grep -v "Autor(a)"|sort|uniq|tr ' ' '_'> Autores.csv
```

O comando anterior Extrai a coluna 1 (Autor(a)) e a coluna 7 (Código) da planilha geral, passa o conteúdo extraído ao **grep** que com o atributo **-v** elimina a linha do título, classifica e elimina as linhas repetidas. O comando **uniq** repassa o resultado ao comando **tr** que substitui os espaços em branco pelo caractere sublinhado (_) e então grava o conteúdo no arquivo Autores.csv.

A substituição dos espaços em branco por sublinhados é realizada para evitar falhas nos comandos executados posteriormente, de forma transitória, e, ao final do processo, será desfeita. Na forma como está escrito o comando **for** mais abaixo, sem a substituição dos espaços em branco, falhariam o próprio comando **for** e o comando **grep**.

```
cut -f1 Autores.csv | sort | uniq > AutoresOrdenados
```

O comando acima extrai somente o nome dos autores (a coluna 1) do arquivo Autores.csv, classifica-os, elimina os duplicados e salva o resultado no arquivo AutoresOrdenados.

```
echo -e "Autor(a)\tPublicações" > AutoresAgrupados.csv
```

O comando anterior cria o arquivo AutoresAgrupados.csv com a linha de cabeçalho do arquivo.

```
for i in $(AutoresOrdenados); do echo -n $i | tr '_' ' '>> AutoresAgrupados.csv; echo -ne "\t" >> AutoresAgrupados.csv; publicacoes=$(grep $i Autores.csv | cut -f2|sort) ; echo $publicacoes | sed 's/ /, /g' >> AutoresAgrupados.csv; done
```

O comando acima é muito similar ao utilizado para a geração do arquivo EspecialidadesAgrupadas.csv. Exceto pelos nomes dos arquivos envolvidos, a mudança substancial é a existência do comando **tr** **'_'''** após o **echo -n $i**. O comando **echo**, de forma similar ao caso da tabela Especialidades, obtém o nome do(a) autor(a) que é repassado ao comando **tr** através do pipe (|) para trocar os caracteres sublinhados por espaço em branco e gravar o nome do(a) autor(a) já correto no arquivo AutoresAgrupados.csv.

Para gerar a Tabela 5 do Apêndice 2 com a relação das Edições vinculadas a eventos foi utilizado o comando abaixo:

```
cut -f10 BDRevistaConscientia.csv | sort | uniq | sed 's/ - /\t/g' | sed -E 's/Vol. ([1-9]) /Vol. 0\1 /g' | sed -E 's/No ([1-4])/No. \1/g' | sort > Edição_e_Evento.csv
```

Foram extraídos os dados da coluna 10 (Dados Complementares), classificados, eliminadas as redundâncias, separados os campos pelo separador hífen (-) alterado para tabulação, acrescentado o zero à esquerda dos Volumes 1 a 9, acrescentado o caractere ponto (.) ao final da expressão **No** onde estava faltando, classificadas novamente porque agora o zero foi adicionado à esquerda, o que muda a ordem de classificação e salvo o resultado no arquivo Edição_e_Evento.csv. Carregada a planilha no LibreOffice Calc, foram eliminadas as linhas em que não haviam eventos, e então os dados da planilha foram copiados e colados no LibreOffice Writer para montar a Tabela 5 do Apêndice 2.

Conclusão

O autor objetivava demonstrar um pouco do que pode ser realizado pelo pesquisador que tenha necessidade de manipular dados armazenados neste formato. Alguns detalhes e particularidades foram necessários ser mostrados para melhor entendimento. Com os comandos apresentados e muitos outros disponíveis, pode-se fazer manipulações de dados muito mais elaboradas, só depende da necessidade e vontade do pesquisador em aprender uma ferramenta adicional ao que conhece.

Produzir as planilhas dos apêndices e mesmo o texto do artigo acabou sendo mais demorado que o pretendido originalmente. A própria manipulação dos dados acabou evidenciando possibilidades que não haviam sido vislumbradas num primeiro momento gerando retrabalhos. A cópia inicial parcial dos dados mostrou ser uma decisão errada, pois posteriormente foi necessário buscar-se dados que não haviam sido copiados, e por ter sido realizadas várias manipulações nos dados foi preciso refazer trabalho já realizado.

Inicialmente, não se tinha ideia de relatar a experiência, então o processo de trabalho não foi adequadamente documentado naquele momento.

O planejamento prévio do resultado final a ser apresentado facilita bastante o processo de manipulação dos dados e criação das tabelas. Neste sentido uma dificuldade foi o fato do autor não ser o "proprietário" dos dados. Por isto a situação ideal é o próprio pesquisador detentor dos dados conhecer os recursos existentes para poder manipular os dados e as ferramentas à sua própria necessidade.

Bibliografia Específica

1. **Vieira,** Waldo; ***Léxico de Ortopensatas;*** revisores Equipe de Revisores do Holociclo; 2 Vols.; 1.800 p.; 652 conceitos analógicos*;* 22 *E-mails;* 19 enus.; 1 esquema da evolução consciencial; 17 fotos; glos. 6.476 termos; 1.811 megapensenes trivocabulares; 1 microbiografia; 20.800 ortopensatas; 2 tabs.; 120 técnicas lexicográficas; 19 *websites;* 28,5 x 22 x 10 cm; enc.; *Associação Internacional Editares;* Foz do Iguaçu, PR; 2014b; página 1.601.

Bibliografia Complementar

1. **Free Software Foundation;** disponível em**<http://www.fsf.org/>;** acesso em: 09.03.2017.

2. **GNU/Linux;** ***Descrição do Sistema Operacional Unix-like;*** disponível em <https://pt.wikipedia.org/wiki/GNU/Linux>; acesso em: 09.03.2017.

3. **GNU Operating System;** disponível em<https://www.gnu.org/>; acesso em: 09.03.2017.

4. **Sed,** a Stream Editor; disponível em <https://www.gnu.org/software/sed/manual/sed.html>; acesso em: 02.04.2017.

5. **Sistema Operacional GNU;** ***Linux and the GNU System;*** disponível em <https://www.gnu.org/gnu/linux-and-gnu.html>; acesso em: 09.03.2017.

6. **Visão Geral do Sistema GNU;** ***Overview of the GNU System;*** disponível em<https://www.gnu.org/gnu/gnu-history.html>; acesso em 09.03.2017.