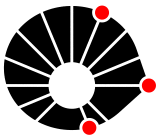

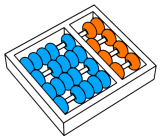

Relatório sobre "Tutorial para Instalação de Repositórios de Dados Científicos: CKAN, DataVerse e DSpace"

Rodrigo Nagamine

Relatório Técnico - IC-PFG-18-33
Projeto Final de Graduação
2018 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Relatório sobre “Tutorial para Instalação de Repositórios de Dados Científicos: CKAN, DataVerse e DSpace”

Rodrigo Nagamine[1]

[1] Instituto de Computação Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6176
13083-970 Campinas-SP, Brasil

ra137531@students.ic.unicamp.br

Resumo. A instalação de ferramentas pode ser complexa devido a sua variabilidade de possíveis configurações. Tendo isso em vista, foi criado um grupo de trabalho de 7 universidades públicas do estado de São Paulo, o qual criou um manual de instalação para 3 ferramentas de repositórios open source, que são: CKAN, DSpace e DataVerse.

O trabalho consistirá na implementação, análise de dificuldades e possibilidades que as ferramentas apresentam, procurando analisar complexidade e permissões de metadados em comum diretamente ou indiretamente possíveis de serem utilizados.

O foco do projeto é fazer uma análise das ferramentas citadas para que facilite na escolha de uma com o intuito de unificar em um repositório, todas publicações acadêmicas de diversas universidades brasileiras.

Este projeto tem como objetivo, fazer uma análise completa do manual de instalação de 3 repositórios e fornecer um feedback para o grupo de trabalho.

Palavras-Chave: Repositórios de Dados Científicos; CKAN; DataVerse; DSpace.

1. Introdução

Considerações Iniciais:

1) Este relatório corresponde ao relatório final da disciplina MC030, orientada pela professora Claudia Bauzer Medeiros.
2) Trata-se de uma iniciativa desenvolvida pelas 7 universidades públicas do Estado (UNICAMP, USP, UNESP, UNIFESP, ITA, UFSCAR, UFABC) e, recentemente, com a adesão do CNPTIA-EMBRAPA. Esta iniciativa está sendo realizada a partir de um Grupo de Trabalho estabelecido pela FAPESP, como parte das iniciativas da Fundação em Open Science e Open Data. Além de aumentar a visibilidade das pesquisas desenvolvidas pelos pesquisadores, a disponibilização dos dados permite novas descobertas e facilita novas cooperações científicas (por meio do reuso e compartilhamento de dados). Não se trata de um único repositório, mas sim de vários repositórios em rede, onde os pesquisadores de cada instituição participante poderão depositar os dados de suas pesquisas. Na verdade, os dados podem ser armazenados em outros locais (por exemplo, na Bioinformática já existem repositórios mundiais); no entanto, o mínimo que será necessário é o armazenamento dos metadados (informações sobre os arquivos de dados de pesquisa), pois a partir deles os dados poderão ser localizados. Cada participante do GT está desenvolvendo os sistemas necessários para este arquivamento e para a gestão confiável de dados de pesquisa.
3) Ao me referir ao PDF fornecido pela professora orientadora Claudia Bauzer Medeiros, estarei utilizando o nome do arquivo, chamado de ManualRepositorios, ou apenas PDF.
4) Frases entre aspas referem-se a comandos ou textos retirados do PDF ou da Wiki o qual é referenciado
5) Resumo das tarefas:
    a) As atividades do projeto se resumem a testar as instruções fornecidas pelo PDF, o qual inclui a wiki de cada repositório.
    b) No caso de não haver problemas, reportar que está correto, e no caso de erro, tentar encontrar alguma solução para que funcione.
    c) Além disso, foram reportados erros ortográficos.
6) Dificuldades encontradas e benefícios de aprendizado:
    a) Devido ao não conhecimento de uso do sistema operacional Ubuntu, houveram dificuldades ao longo dos testes, dificultando o entendimento de bugs, e por isso, dificuldade com debugs de algumas instruções fornecidas.

b) Por outro lado, haver várias dificuldades resultou em um grande aprendizado ao realizar esse projeto. Aprendizado com o uso do sistema operacional Ubuntu, uso de ferramentas do mesmo.

7) O objetivo desse projeto foi fazer uma análise completa do manual e fornecer feedback para o grupo de trabalho.

Resumo:

A instalação de ferramentas pode ser complexa devido a sua variabilidade de possíveis configurações. Tendo isso em vista, foi criado um grupo de trabalho de 7 universidades públicas do estado de São Paulo, o qual criou um manual de instalação para 3 ferramentas de repositórios open source, que são: CKAN, DSpace e DataVerse.

O trabalho consistirá na implementação, análise de dificuldades e possibilidades que as ferramentas apresentam, procurando analisar complexidade e permissões de metadados em comum diretamente ou indiretamente possíveis de serem utilizados.

O foco do projeto é fazer uma análise das ferramentas citadas para que facilite na escolha de uma com o intuito de unificar em um repositório, todas publicações acadêmicas de diversas universidades brasileiras.

Este projeto tem como objetivo, fazer uma análise completa do manual de instalação de 3 repositórios e fornecer um feedback para o grupo de trabalho.

2. Ckan:

Não está claro no começo do manual sobre o que será feito. Inicialmente o manual em PDF diz para realizar a instalação do CKAN utilizando o manual que está na wiki, que foi baseada em um manual que está no site oficial do CKAN. Ficou bem confuso qual dos dois manuais deve-se seguir, se é o do site oficial ou o da wiki, para começar a instalação. Depois de observar ambos links, percebe-se que o site oficial possui a instalação em inglês, e a wiki a instalação em português. É possível observar também que as configurações de pastas e permissões são diferentes entre ambas, e para melhor continuidade, deve-se seguir as instruções fornecidas pela wiki, apesar de não estar explícito que não é necessário instalar o CKAN pelos links fornecidos anteriormente.

No ManualRepositorios: em 1.3.1 há passos intermediários durante a instalação e configuração inicial da wiki. Ou seja, não se deve seguir a instrução 1.3 as cegas e realizar toda instalação fornecida na wiki, sem antes terminar de ler os próximos passos, pois em 1.3.1 há comandos para serem rodados entre as etapas da wiki.

Na wiki (IBICT) não é especificado como que se faz para editar arquivo (não intuitivo para leigos). Poderia haver alguma recomendação por parte do manual, antes do ponto 1.3, para que sempre que houver algum comando de edição de texto, utilizar um comando simples de

editor de texto, como “gedit [nome do arquivo]”, como acontece no passo “Crie o arquivo de configuração do CKAN: Crie o arquivo de configuração do CKAN: Agora edite o arquivo criado para configurar a conexão com o banco de dados PostgreSQL.”

Na parte de configuração de Jetty-Solr:

“Edite o arquivo '/etc/default/jetty8' e modifique as seguintes variáveis:”

Ou seja, pede-se para editar o arquivo, ou seja, utilizar algum comando como “gedit” ou “nano”, é necessário utilizar “sudo”, e um não funcionou para “jetty8”, e sim para “jetty”. Então o comando executado deve ser “sudo nano /etc/default/jetty”.

Observação: dentro da wiki, toda vez que citar o arquivo “Jetty8”, deve ser utilizado o nome “Jetty”.

Na linha “Abra o seu navegador e entre em: 'http://127.0.0.1:5000/'; Você será capaz de ver o CKAN em funcionamento já.”, ao clicar no link, dá-se erro, pois há aspas simples no final desse hiperlink. Deve-se utilizar o link “'http://127.0.0.1:5000/;

Configurar o Apache e o Nginx no CKAN:

Não houve problemas

Configurar o DataStore

Na linha:

“Obs.: Modifique a palavra 'pass' para as respectivas senhas criadas. “

Ao ler “respectivas senhas criadas”, parece que foram criadas mais de uma senha, mas na verdade foi criado apenas uma senha e é para utilizá-la nos dois campos que forem pedidos.

Configure o FileStore

Sem problemas

Configure o DataPusher

Sem problemas

Nota: Você será capaz de ver o CKAN sendo executado em 'http://127.0.0.1:8080/' (Apache) e 'http://127.0.0.1/' (Nginx).

Fim da parte da wiki.

2.3.1. Passo complementar e adição e administrador

Sem problemas

2.3.2. Permissões de usuário no CKAN:

Sem problemas

2.3.3. Extração de dados pelo Metabuscador

Não ficou claro se é necessário fazer algo. Ao ler o texto, entende-se que o texto está apenas explicando sobre a extração de dados pelo Metabuscador. Ou seja, é uma nova sessão no manual, que é um trecho do manual apenas informativo e nada é feito no ponto de vista de instalação.

Caso o usuário que está configurando o ckan queira conhecer mais sobre isso, o material para isso é fornecido.

2.3.4. Habilitação do Protocolo OAI-PMH

Na parte de: “É preciso instalar as seguintes dependências para que o protocolo funcione:”

Não ficou claro que é necessário estar no ambiente do ckan (python) para as instalações.

Além disso, não há muita explicação das dependências e de possíveis erros em suas instalações pelo git. No caso, não foi encontrado o arquivo “lber.h” e “sasl/sasl.h”.

Python ldap necessita dessas bibliotecas, sendo então necessário instalá-las com

sudo apt-get install libldap2-dev

sudo apt-get install libsasl2-dev

Na verdade, esses comandos estão escritos para executar depois dos links para instalação das dependências. Poderiam e deveriam estar antes de fornecer os links do git para evitar erros.

Durante a instalação de uma terceira dependência exigida durante esse passo, de acordo com as documentações no git, há uma parte que diz ser opcional.

“Install ckanext-harvest (https://github.com/ckan/ckanext-harvest#installation) (Only if you want to use the RDF harvester)”

No caso, não é opcional, pois após continuar para a parte dos termos, notei que o ckan não estava funcionando corretamente, principalmente devido a não instalação dessa dependência. Acredito que ter um aviso no PDF evitaria esse erro e/ou dúvida.

Após instalá-la, não é dito que é necessário iniciar o redis. Para isso, é necessário utilizar o comando:

Sudo service redis-server restart.

Necessário também reiniciar o apache para voltar a funcionar.

Sudo service apache2 restart.

“2.3.5. Adilção do termo de uso do repositório CKAN”

Erro ortográfico no texto. O correto deveria ser Adição.

Na frase:

“Para se inserir o link **para** do termo de uso em pdf na tela de registro foi preciso alterar o arquivo:” - Essa frase também apresenta erros de português. A forma que acredito ser a correta deveria ser “Para inserir o link do termo de uso em pdf na tela…”

2.3.6. Alteração de tradução de Mantenedor para Financiador

No tópico 6, não ficou claro como fazer para chegar até a imagem desejada para comparação.

Depois de tentativas, foi descoberto que deve estar logado no sistema para ter acesso a tela desejada, e com o usuário “default” (usuário “comum criado” não funcionou para poder comparar), o qual não me recordo de ter definido uma senha para o mesmo (tentei uma senha padrão que utilizei para testes, mas não obtive sucesso). Para isso, foi alterado via terminal a senha do usuário default e então pude encontrar os campos para comparação e verifiquei que eles foram alterados com sucesso.

No tópico número 7, a frase está “7 - A tela com informações **adicionar**” - Acredito que o correto seria “**adicionais**”

Já a figura 3, não consegui encontrar como chegar nela para poder comparar.

2.3.7. Cadastro de um novo conjunto de dados

Telas verificadas e compatíveis com as do manual

2.3.8. Autorização dos usuários no CKAN

Telas verificadas e compatíveis com as do manual.

3. Dspace

3.1. Instalação de pacotes necessários

“apt-get install postgresql openjdk-8-jdk-headless maven ant apache2”

Faltou informar que seria necessário realizar sudo no começo. Apesar de ser dito para executar como usuário root, acredito que seja melhor na linha do comando deixar especificado para utilizar sudo (para melhor clareza para usuários leigos), e se o usuário já estiver como root, não haverá problema.

3.2.Criação de usuário do sistema operacional

Sem problemas.

3.3.Instalação do Dspace

Sem problemas.

Dentro da wiki:

Database Setup

“createuser --username=postgres --no-superuser --pwprompt dspace”

Deu erro nesse comando por não conseguir se comunicar com postgres. Para executar o comando, foi necessário adicionar antes “sudo -u postgres”.

Esse erro é o que é avisado por "superuser" account (e.g. postgres)”

Para os comandos:

“createdb --username=postgres --owner=dspace --encoding=UNICODE dspace”

“psql --username=postgres dspace -c "CREATE EXTENSION pgcrypto;"

O mesmo erro acontecia para os comandos acima, os quais também foram solucionados ao adicionar no início deles “sudo -u postgres”.

Wiki: 5.Initial Configuration (local.cfg):

Não está nem um pouco claro o que deve ser feito para copiar o arquivo de configuração de exemplo que ele sugere.

Para isso, é necessário navegar até a pasta instalada por meio do comando “cd”, e utilizar o comando “cp local.cfg.EXAMPLE local.cfg” (copy).

Aparentemente a única modificação necessária para fazer posteriormente é o diretório (dspace.dir), como especificado no site, ou seja, para a mesma pasta em que foi instalado o DSpace.

Wiki: 6.DSpace Directory:

Sem problemas.

Wiki: 7.Build the Installation Package

Para os comandos:

“cd [dspace-source]” - Não é tão intuitivo como os outros, pois deve-se navegar até a pasta source do dspace, e não apenas utilizar o comando dado.

“mvn package” - Precisa instalar pacote antes com Sudo apt-get install maven, e o comando “mvn package” demora por volta de 20 minutos, poderia haver um aviso para melhor planejamento.

Wiki: 8.Install DSpace

O comando “ant fresh_install” - deu erro.

Inicialmente houve erro de login, pois não estava rodando como usuário DSpace, e sim com meu usuário padrão. Para isso foi necessário “sudo su DSpace” para resolver esse problema.

Depois disso, erro com pacote, foi necessário reinstalar o “mvn package”, que levou mais 20 minutos.

Ainda assim, houve erro de versão do postgres e da extensão pgcrypto. (9.3 e 1.0, e precisam ser respectivamente 9.4 e 1.1)

Para isso foram rodados os comandos

sudo add-apt-repository "deb https://apt.postgresql.org/pub/repos/apt/ trusty-pgdg main"

wget --quiet -O - https://postgresql.org/media/keys/ACCC4CF8.asc | sudo apt-key add -

sudo apt-get update

sudo apt-get install postgresql-9.4

Mesmo após estes comandos, a instalação do postgres 9.4 resultou em erro para instalá-lo, pois ele inicializa os dois e tenta utilizar o 9.3.

Tentei desinstalar o 9.3, mas mesmo desinstalando, ao rodar o comando de instalar o DSpace "ant fresh_install", o erro ainda permanecia por problemas de versão.

Descobri que é possível executar apenas o 9.4, mas mesmo assim, o erro persistiu, provavelmente por haver algum arquivo de configuração setando para a versão 9.3.

Foi rodado tudo mais uma vez, porém agora mudando os arquivos para usarem a porta 5433, para então esses arquivos utilizarem o postgres 9.4.

A instalação com "ant fresh_install" deu certo.

Wiki: 9. Decide which DSpace Web Applications you want to install.

Não é fornecido muita informação, para leigos, qual seria o melhor Web Applications, ou qual seria o Web Application mais adequado (ou suficientemente bom) para prosseguir.

Além disso, no passo "**Wiki: 10. Deploy Web Applications:** Please note that in the first instance you should refer to the appropriate documentation for your Web Server of choice. The following instructions are meant as a handy guide. You have two choices or techniques for having Tomcat/Jetty/Resin serve up your web applications:" ficou muito vago sobre o que deve ser feito. Não é explicado como instalar uma web application e nem há as pastas de webapps para rodas os comandos:

"cp -R [dspace]/webapps/* [tomcat]/webapps* (This will copy all the web applications to Tomcat).

cp -R [dspace]/webapps/jspui [tomcat]/webapps* (This will copy only the jspui web application to Tomcat.)"

Foi escolhido e copiado o webapp "xmlui" para dentro da pasta do webapps do jetty. Este foi o webapp escolhido pois pela descrição parecia o melhor considerando simplicidade e atender os requisitos desejados.

As dificuldades encontradas foi de ter que encontrar aonde se encontravam as pastas Webapps de ambos servidores web.

Wiki: 11.Administrator Account:

O comando "[dspace]/bin/dspace create-administrator" deixa ambíguo qual pasta bin deve ser usada. Tentei utilizar a basta dentro "dspace/bin", mas deu erro. O correto foi utilizar a pasta que vem do zip, no caso "dspace-6.3-realeses"

Wiki: 12.Initial Startup!

http://localhost:8983/xmlui/ está funcionando sem problemas. Conseguir logar no usuário de administrador criado.

Fim da wiki do DSpace

3.4. Configuração pós instalação

a) Metadados personalizados

Sem problemas para as configurações pós instalação.

Aparentemente, há imagens prontas para o dspace, que seria necessário rodar apenas “docker pull” e “docker run”.

4. DataVerse

Foi utilizado a versão ubuntu 14.04, pois no PDF é dito que ele foi baseado nessa versão de Linux.

4.1 e 4.2 são apenas informativos. Ok.

4.3 - Não está claro que é para seguir o manual disponibilizado no link fornecido. Aparenta ser apenas informativo com links de origem, como nos dois passos anteriores.

O manual disponibilizado no link, é bem mais técnico e menos intuitivo, comparado a wiki dos outros dois repositórios anteriores.

Não há um passo a passo bem explicado sobre o que deve ser feito e como deve ser feito. Acredito que no manual em PDF poderia ter uma breve explicação do que tem no link.

Por exemplo:

"Há a sessão de preparação que informa sobre requisitos necessários e algumas decisões a serem tomadas, como por exemplo, escolher um instalador.

Depois há uma sessão de pré-requisitos e de como instalá-los."

No caso, as instalações necessárias de pré-requisitos que tive que fazer, são:

-GlassFish

Durante a instalação, ocorreu tudo normal, exceto na parte em que é pedido para iniciá-lo e verificar a versão: "Start Glassfish and verify the Weld version:

/usr/local/glassfish4/bin/asadmin start-domain"

Ao executar o comando, o fica-se no aguardo do domínio para ser inicializado, mas nada acontece por muitos minutos. Fica a mensagem "Waiting for domain1 to start……"

-JQ

Instalação sem nenhum problema

-R

Os comandos fornecidos para instalação mínima do R para que funcione, não funcionam.

É utilizado o comando "yum" para instalá-los, porém estes não funcionam no Ubuntu, o qual é o sistema operacional utilizado, como descrito no início. E como descrito

no manual, o R é o responsável para lidar com todos os arquivos de dados tabelados que serão utilizados, tornando a instalação incompleta e não utilizável

-ImageMagick

Também utiliza “yum” para instalação.

A inconsistência encontrada é que no manual em PDF, logo no início da instalação do Dataverse é dito que o manual foi baseado para a versão 14.04 do Ubuntu, e também baseado no manual de instalação do link fornecido http://guides.dataverse.org/en/latest/installation/. Porém este último utiliza comandos para RHEL/CentOS 6 ou 7 e não para o Ubuntu

Na parte de instalação, há um campo escrito **“Important”** dentro de **“Running the Dataverse Installer”**, o qual explica sobre a importância de colocar o executável no PATH, mas não explica como fazê-lo, e que se o instalador não se sente confortável com o que está sendo falado, é melhor não começar, pois é provável que não funcionará.

Após não conseguir instalar 2 pré-requisitos, não foi dada continuidade a testes de instalação, pois não seria possível saber se o problema é com o manual ou a falta de pré-requisitos.

5. Conclusão:

Como conclusão após seguir o manual em PDF, ele está detalhado e uma boa base para aqueles que entendem o básico de computação para prosseguir com a instalação dos repositórios, mesmo que não perfeito, como apontado em pontos de erros neste relatório. Porém para leigos, há pontos de possíveis melhorias com explicações ainda mais claras do que pode e precisa ser feito. Acredito que se um usuário está designado a este tipo de tarefa, ele possuirá um conhecimento mínimo e não encontrará problemas básicos, como os que eu encontrei.

Por outro lado, esse projeto me permitiu utilizar um sistema operacional que poucas vezes tive a necessidade de utilizar durante a graduação, mesmo que muitas vezes recomendado, as instruções eram sempre claras e simples. Dessa vez tive que conhecer e aprender mais por conta própria para entender os problemas e tentar resolvê-los.

Além disso, o tempo gasto para toda configuração e comandos rodados, poderia ser substituído com a utilização de dockers já prontos. Estes estão bem explicados de como rodá-los e quais seriam os comandos rodados por eles, para que suas configurações básicas sejam realizadas, além de permitir futuras configurações pessoais que atendam às necessidades do instalador.

Seguem os links de imagens de Dockers que poderiam ser salvos e guardados caso necessitados futuramente, para Ckan e o Dataverse respectivamente.

CKAN: https://hub.docker.com/r/ckan/ckan/

Dataverse: https://hub.docker.com/r/ndslabs/dataverse/

Para o DSpace bastaria seguir os seguintes comandos

```
# criar rede para comunicação entre containers
docker network create rede
# ligar o banco
docker run  --network=rede --name=dspacedb  dspace/dspace-postgres-pgcrypto
# ligar dpsace
docker run --network=rede --name=dspace -p 127.0.0.1:8080:8080 -p 127.0.0.1:8009:8009 dspace/dspace:dspace-6.3
# criar usuario de administrador
docker exec -it dspace /dspace/bin/dspace create-administrator
```

Referências

Arquivo fornecido pela professora orientadora Claudia Bauzer Medeiros, com o nome ManualRepositorios, cujo título do material é “Tutorial para Instalação de Repositórios de Dados Científicos: CKAN, DataVerse e DSpace”.

Manual de instalação CKAN http://wiki.ibict.br/index.php/Manual_de_instalação.

Manual de instalação DSpace https://wiki.duraspace.org/display/DSDOC6x

Manual de instalação DataVerse http://guides.dataverse.org/en/latest/installation/