PROPOSTA DE ADOÇÃO DE SOLUÇÃO DE SOFTWARE COM CÓDIGO ABERTO PARA INDEXAÇÃO DE ARQUIVOS DE TEXTO

Fernando Prudêncio de Souza

Resumo: Esta pesquisa trata do atendimento de demandas de arquivamento e busca de altos volumes de documentos em formato *PDF* (portable document format), seja no modo texto nativo ou imagem, que tendem a consumir recursos consideráveis da PJC/MT (Polícia Civil do Estado de Mato Grosso). Através de um comparativo entre a ferramenta selecionada e os principais concorrentes do mercado, apresentaremos as vantagens e desvantagens da solução de maneira a prover as informações necessárias para uma tomada de decisão.

1 Introdução

Esta proposta tem por objetivo a apresentação de um sistema *GED* (Gestão Eletrônica de Documentos), para adoção e integração com os demais sistemas e processos utilizados pela PJC/MT.

As possibilidades de integração são baseadas nos conteúdos expostos durante o processo seletivo nº 001/2020/PJC/MT. Não é possível afirmar que os softwares mencionados façam parte da estrutura interna de trabalho da PJC/MT.

Na proposta, serão analisados três softwares de código livre. Dois deles são classificados como generalistas, voltados aos processos de *GCE* (Gerenciamento de Conteúdo Empresarial), os quais englobam, entre outros processos, o *GED*: Nuxeo e Alfresco. O último, especialista, é uma solução direcionada em *GED*, requisito essencial para solução da problemática proposta.

2 Comparativo da Soluções

2.1 Breve Apresentação

2.1.1 Alfresco ECM

Alfresco, segundo sua página oficial, é uma empresa norte americana de software de código aberto fundada em 2005, com sede em Boston, Massachusetts, EUA.

O sistema que leva o nome da empresa, tratava-se inicialmente de um *GED* e hoje se subdivide em uma gama de sistemas para composição de um *GCE*.

Entre seus cases de sucesso se encontram nomes conhecidos, como a NASA (agência espacial americana) e CISCO (tecnologia da informação e redes).

2.1.2 Nuxeo ECM

Conforme relatado em seu perfil oficial no LinkedIn, é uma empresa norte americana de software de código aberto fundada em 2008, com sede em New York, New York, EUA.

Assim como Alfresco, tratava-se de um sistema GED e hoje abrange uma gama de sistemas CGE.

Em sua lista de clientes, destacam-se os nomes LOREAL (indústria cosmética), EA (jogos eletrônicos), FOX (entretenimento), entre outros.

2.1.3 Papermerge

Fundada em 2017, com sede em Berlin, Alemanha, é descrita em sua página oficial como um software *GED* que teve início para solução de um problema específico na organização de documentos digitalizados; evoluindo de forma orgânica para o formato que possui hoje.

Inicialmente de código proprietário, migrou para modalidade de código livre em 06/01/2020 (seis de janeiro de dois mil e vinte).

Até o presente momento, a empresa não divulgou sua lista de cases.

2.1.4 Comparativo

	Papermerge	AlfrescoE CM	Nuxeos ECM
OCR	SIM	SIM	SIM
Metadata	SIM	SIM	SIM
Diretórios	SIM	SIM	SIM
Permissões por avançadas por Usuário/Documento	SIM	SIM	SIM
Versionamento Documentos	NÃO	SIM	SIM
Assinatura Digital	NÃO	SIM	SIM
Workflows	NÃO	SIM	SIM
Tags	SIM	SIM	SIM
REST API	SIM	SIM	SIM

2.1.5 Seleção

Embora as opções Alfresco e Nuxeos apresente maiores opções, ambos tratam de uma gama maior de recursos GCE, o que exige maior suporte e equipe de manutenção, além de uma maior curva de aprendizagem.

A opção restante se trata da solução específica para o uso de *GED*. Por se tratar de software especialista, o Papermerge atende a todos os requisitos apresentados na avaliação, com uma menor curva de aprendizagem em comparativo com as demais soluções apresentadas sendo, portanto, a opção selecionada.

3 Implantação

3.1 Recursos Necessários

3.1.1 Recursos de Software

Softwares adicionais necessários para a execução do servidor:

- **Python**>= 3.7 – A aplicação é escrita em Python, sendo o mesmo é necessário para execução do servidor;
- **Django**>= 3.1 – É utilizado como biblioteca de componentes web;
- **Tesseract** – OCR de código livre da Google;
- **Imagemagick** – Para conversão de imagens em formatos diversos;
- **Poppler** – Para operações/manipulações de PDFs.

3.1.2 Recursos de Hardware

O Papermerge pode ser configurado em uma única instância ou de maneira distribuída, sendo um servidor para processamento e outro para componentes web.

Configuração mínima de hardware recomendada:

- Servidor Web
  - 900 MHz CPU
  - 512 MB RAM
  - 15 GB de espaço em disco;
- Servidor para Processamento de Arquivos
  - 1 GHz CPU;
  - 1 GB RAM;
  - 25 GB de espaço em disco;
- Servidor único (Web + Arquivos)
  - 1 GHz CPU;
  - 1 GB RAM;
  - 25 GB de espaço em disco;

3.2 Etapas de Implantação

3.2.1 Instalação

Embora a documentação oficial forneça detalhes sobre a instalação em servidores físicos, a proposta apresentada adota a linha retratada na avaliação e faz uso da instalação via docker.

Existem duas imagens homologadas pelo Papermerge para implantação: *LinuxServer.io Image* e sua imagem oficial *Eugenci*

Papermerge. Pela maior clareza na documentação, foi selecionada a imagem LinxuServer.io.

Configuração relevante:

- image: ghcr.io/linuxserver/papermerge:version-v1.5.5
- environment
  - PUID=1000
    - Id de usuário padrão para configuração de permissões a diretório;
  - PGID=1000
    - Id de grupo de usuários padrão para configuração de permissões a diretório;
  - TZ=America/Cuiaba
    - Configuração de fuso horário do servidor;
- volumes:./ecm/conf:/config
  - Para customização do arquivo de configurações, em especial, alterar o acesso padrão de banco de dados SQLite para PostgreSQL.

Uma instalação via docker está disponível no repositório desta mesma avaliação.

3.2.2 Migração

A migração de arquivos existentes pode ser realizada através da execução de script de importação, localizado na pasta raiz da instalação (manage.py), onde todos os arquivos da pasta informada como parâmetro serão importados para dentro do servidor.

Opcionalmente, o upload de arquivos pode ser realizado via REST API. Para integrar uma aplicação, é necessário gerar um token através do Papermerge e, de posse deste token, consumir as APIs de upload de arquivos.

Não há forma documentada de migração entre bancos de dados, mas é uma possibilidade a ser estudada caso o volume de arquivos legados exceda as capacidades das Apps desenvolvidas.

4 Resultados Esperados

Com a eficiência de um sistema focado apenas em *GED*, integrado as aplicações existentes e potencialmente inserido no contexto de banco de dados já implantado; espera-se com o Papermerge:

- Uma solução de alta aderência e baixa curva de aprendizado, possibilitando o armazenamento e recuperação eficiente de dados binários;
- A redução significativa no consumo de recursos humanos da PJC/MT que hoje é dedicado ao processo de armazenamento e recuperação destes dados.