RECUPERAÇÃO E CONSOLIDAÇÃO DO ACERVO DE DADOS GEOQUÍMICOS DA CPRM: UMA EXPERIÊNCIA *FREE AND OPEN-SOURCE* COM MÓDULOS PYTHON (GEO)DJANGO, PANDAS E MATPLOTLIB

Mota, C. E.[1]; Alvear, M.[1,2];

[1]CPRM/Serviço Geológico do Brasil, ERJ, Rio de Janeiro-RJ; [2]Universidade Federal do Rio de Janeiro - Dep. Geografia;

RESUMO: A CPRM possui um valioso acervo de informações geoquímicas, que, desde a década de 70, que engloba cerca de 450.000 amostras geoquímicas de diversos materiais geológicos, analisadas por diversos métodos analíticos. Atualmente estes dados estão distribuidos em variadas fontes, com características distintas. Um dos desafios da Divisão de Geoquímica (DIGEOQ) é consolidar todas as informações disponíveis na CPRM em um único acervo, para fins de planejamento, gestão, validação e disponibilização no GEOBANK. O uso de software livre é incentivado como política de governo e o objetivo deste trabalho é apresentar o "*modus operandi*" da consolidação do acervo, como estudo de caso da utilização de softwares de domínio público, desenvolvidos na linguagem de programação Python. O uso de Python é justificado por variadas razões, onde destaca-se a simplicidade e legibilidade do código-fonte, extensa documentação, inúmeras bibliotecas (*desktop*, *web*, *data analysis*, etc) e por ser multiplataforma (Windows, Linux e Mac) e integrada a outros sistemas. Para este caso, foram utilizados as bibliotecas Django 1.9 (*framework web* e *scripts*), Pandas 0.17.1 (*data-analysis* e estatística) e Matplotlib 1.2.0 (representação gráfica 2D) e os dados, armazenados em banco de dados PostgreSQL 9.3/PostGIS 2.0.7. O modelo de dados apresenta a definição de visita de campo, com atributos temporal (data de visita) e espacial (coordenadas numéricas primárias nos sistemas originais e geometrias reprojetadas em SIRGAS2000), que contém uma ou mais amostras e a representação do resultado analítico como ternário amostra-analito-método analítico. As definições descritas no modelo seguem as diretrizes dos manuais geoquímicos da CPRM. As classes/objetos (visitas, amostras, resultados) foram implementadas em modelos no software Django. As conversões das coordenadas para objetos *geometry* foram feitas no próprio modelo, sem a utilização de *triggers* no banco de dados. A partir das definições dos modelos e das tabelas Pandas, foram desenvolvidos vários *scripts*, com requisitos mínimos de informações, para a carga dos dados geográficos e dos boletins analíticos. Por fim, um protótipo de aplicação web Django+Bootstrap, com controle de acesso, foi desenvolvido para tornar amigável a execução dos *scripts* de alimentação pelos colaboradores da DIGEOQ e visualização dos dados carregados. O desenvolvimento de toda a estrutura foi de forma rápida e simples (em poucos dias), conforme a filosofia da linguagem, pois o próprio Django gera as definições do banco de dados e, aliado ao Pandas, detém a representação da inteligência do sistema. Os módulos de administrativo e de autorização internos do Django, permitem manipular as informações, com registro da identificação de usuários e de horarios. A utilização do Pandas e Matplotlib para processamento e representação diminuiu significamente o tempo de desenvolvimento dos *scripts*, pois simplificou a realização de operações tabulares, como junções de resultados, estatísticas básicas e *pivot*, além da representação gráfica. A simplicidade da linguagem Python e as bibliotecas Django, Pandas e Matplotlib, em suma, constituem ferramentas poderosas não só para a construção de complexos sistemas geoespaciais, mas também para elaboração de *scripts* de manipulação de dados em poucas e compreensíveis linhas de código.

PALAVRAS-CHAVE: DATA ANALYSIS, SOFTWARE LIVRE, PYTHON