Relatório Individual Sumarizado para bolsa Fapesp de Treinamento Técnico TT-4A

23/02/2013

Título *do Projeto*:

Desenvolvimento de um sistema de informações integrado para pesquisas sobre políticas públicas. Parte do projeto temático "*Brasil, 25 anos de democracia – balanço crítico: políticas públicas, instituições, sociedade civil e cultura política*" desenvolvido pelo Núcleo de Pesquisas em Políticas Públicas da USP, o NUPPs.

Nome do bolsista:

Flávio Sant Ana Daher

Nível e período de usufruto da bolsa:

Trata-se de uma bolsa de *Treinamento Técnico nível 4-A*, dentro do período de *seis meses*, de 1º de agosto de 2012 até 31 de janeiro de 2013.

Descrição sumarizada das atividades do bolsista:

A contribuição do bolsista da categoria de Treinamento Técnico é contribuir através de sua vivência e experiência na área da Tecnologia de Informação com o projeto de pesquisa acadêmico.

Em linhas gerais, o trabalho desenvolvido nesta etapa foi o da identificação das necessidades particulares de cada uma das diferentes áreas que compreendem uma pesquisa em gestão pública e o do mapeamento das principais demandas por informação da rede de pesquisadores. O objetivo principal, no plano a médio prazo (em dois anos) é o do desenvolvimento de uma solução computacional que permita que os pesquisadores e demais envolvidos na produção da pesquisa sejam abastecidos por dados estatísticos confiáveis e precisos de uma maneira prática e direta, sem haver necessidade de lançarem mão de pacotes ou técnicas de manipulação de dados.

Mais especificamente, neste primeiro momento foi eleito um rol suficiente de fontes de informação primária - bases e microdados disponíveis na internet ou em pacotes particulares -, a partir do qual será possível criar um abrangente banco de dados centralizado, sobre o qual será possível analisar o panorama das diferentes áreas da pesquisa, sempre utilizando uma perspectiva histórica. As informações foram organizadas de modo a ser composta uma base de dados unificada e central de modo a servir como alimento de um sistema interativo de extração, tabulação e cálculo de estatísticas, operado diretamente pelos pesquisadores, e que essencialmente deverá ser intuitivo, de modo a promover insights por conta do maior foco nas análises em si.

Além das atividades técnicas e computacionais realizadas no Núcleo, o bolsista também teve a oportunidade de participará de todo o processo de discussão conceitual com os pesquisadores do projeto **Brasil, 25 anos de democracia – balanço crítico: políticas públicas, instituições, sociedade civil e cultura política** no Núcleo de Pesquisas de Políticas Públicas da Universidade de São Paulo, sobretudo no que se tange à viabilidade de análises quantitativas, e ainda promoveu alguns treinamentos das ferramentas implementadas para os demais colegas.

Detalhamento técnico das atividades do bolsista:

1) Mapeamento das demandas por informações no Núcleo

O projeto temático principal, o qual este auxílio de capacitação técnica visa atender, se subdivide basicamente em cinco áreas dentro do assunto Políticas Públicas. Cada uma destas áreas tem um foco específico e por conseguinte demandas particulares em relação à obtenção, organização e disponibilização de informações.

Como atividade inaugural, o bolsista trabalhou diretamente com os representantes (a saber, os responsáveis, pesquisadores e auxiliares) de cada área, em rodadas de reuniões e de consultas técnicas, visando aprender mais sobre os temas, e, principalmente, mapear as necessidades específicas da pesquisa e eventuais demandas para auxílio computacional para o trabalho colaborativo entre os membros dos grupos.

Em suma, as demandas de cada área são as seguintes:

a) Instituições e Representação Política: Obtenção de bases de dados do TSE com resultados de eleições passadas e perfis de candidatos e eleitos. Obtenção, de forma automatizada, de dados sobre Projetos de Leis, disponíveis nas páginas das instituições (Câmara, Senado, etc). O grupo ainda conta com um banco de dados particular de sobre determinados projetos de leis e de indivíduos relevantes para a pesquisa, que poderiam ser sincronizados com o banco de dados central.

b) Sociedade Civil e Cultura Política: Obtenção de dados socioeconômicos da população brasileira, a partir principalmente das pesquisas amostrais como a PNAD e do Censo Populacional. O Grupo ainda conta com uma série histórica de resultado de pesquisas particulares (*surveys*) que avaliam a cultura política na população brasileira. Fica também a ideia de integrar estes resultados com o banco de dados central.

c) Políticas Públicas de Educação: Obtenção de bases de dados do INEP e MEC, principalmente as séries históricas dos censos do ensino básico e do ensino superior. Outras bases de dados serão também consideradas, como as de avaliação de alunos e de instituições de ensino (ENEM, ENADE, PROVÃO, SAEB), censos e avaliação da Pós-Graduação (CAPES), e censo dos docentes e servidores do ensino superior.

d) Políticas Públicas de Segurança: Obtenção de bases de dados de censos demográficos (já mencionados) e acompanhamento de estatísticas de saúde (DATASUS, Fundação SEADE).

e) Políticas Públicas de Cultura: Identificação de instituições culturais e seus dados atualizados e indivíduos envolvidos, a partir de pesquisas nas páginas das próprias instituições e de suas mantenedoras. O grupo ainda conta com um banco de dados particular com a coleta destas informações, proveitoso caso fosse possível integrá-lo com o banco de dados central.

Após avaliação das demandas dos pesquisadores, escolheu-se como prioridade a montagem do banco de dados central, a partir da compilação de dados das fontes mencionadas. Para tanto, foi-se necessário o desenvolvimento de um ambiente de TI que pudesse receber e hospedar de forma segura e confiável, um sistema de informações. A próxima seção trata deste ponto.

2) Implantação de infraestrutura de TI

Encontravam-se disponíveis no NUPPs duas máquinas do tipo servidores, de porte razoável. Optou-se por colocar em operação dois sistemas em paralelo, um baseado no Windows Server (acompanhando a

infraestrutura de rede que já estava em operação) e outro tendo como base o Linux, com objetivo de servir como ambiente de desenvolvimento das tecnologias.

De forma a atender as demandas, procurou-se desenvolver os ambientes tornando possível a implantação de uma solução que integra a organização de dados em bancos com a análise estatística, um conceito conhecido como OLAP, a qual torna possível o cruzamento entre diferentes dados por meio de dimensões comuns e ainda dá suporte à interação com os dados através das hierarquias e das possíveis operações OLAP. Foi usado para este propósito o pacote de código-aberto *Pentaho*.

As máquinas atualmente estão configuradas da seguinte forma:

- Servidor Windows Server (para produção), com SQL Server e ambiente de programação Visual Studio.
- Servidor Linux (para desenvolvimento) com Apache, PHP, MySQL, Tomcat, Mondrian, Python, R, entre outros aplicativos.
- Em ambos os servidores: Uma suíte de BI (Pentaho), um sistema do tipo OLAP, capaz de transformar tabelas relacionais em multidimensionais, tornando possível a capacidade de interagir com os dados, por hierarquia.
- Banco de Dados com mais de 10GB de tabelas padronizadas contendo dados sobre ensino superior e demais dados sociodemográficos de todo país, em sequência histórica, de 1998 até 2010 (detalhadas a seguir).

3) Bases de dados e indicadores

Como proposto, logo após a identificação das demandas por informação por parte dos pesquisadores do Núcleo, foi obtida as fontes de dados com as cargas mais recente disponíveis a partir dos endereços eletrônicos das organizações responsáveis e por meio de solicitação por meio de contatos com os departamentos de divulgação de estatísticas.

Os dados recebidos passaram por um longo processo de limpeza e padronização, de modo a se adaptarem à estrutura usada no repositório atual (o sistema OLAP), para permitir a integração e intercomunicação das bases. O diagrama de tabelas e relações correspondentes à arquitetura do banco de dados central segue a seguir.

A seguir segue uma tabela com a relação das fontes escolhidas para a primeira etapa de carregamento do sistema de informações, já no formato de *cubos OLAP*:

Tabela (cubo)	**Fonte**
ProUni	SISPROUNI
Docentes	INEP - Censo da Educação Superior
Servidores	INEP - Censo da Educação Superior
Enade	INEP - Resultados do ENADE
Capes	CAPES - Sistema Geocapes
Educação Superior	INEP - Censo da Educação Superior
População	SEADE (São Paulo) e IBGE (Brasil)
Ensino Médio	INEP - Censo da Educação Básica
EAD	INEP - Censo da Educação Superior

4) Demais tarefas, conforme o plano de atividades

Como atividades desenvolvidas pelo bolsista, constam:

- Implementação de rotinas e sistema de Mineração de Dados: a partir da inclusão das bases de dados já padronizadas no sistema de banco de dados SQL Server (Windows), foi possível iniciar os testes com o componente nativo Analysis Services, um pacote que dá suporte a atividades de Mineração de Dados, além de oferecer também um ambiente OLAP alternativo. Apesar dos testes iniciais, o sistema de Mineração de Dados está programado para ser usado nas próximas etapas, quando o foco dos trabalhos forem justamente as análises.

- Construção de aplicativos para coleta automática de dados a partir da web: Algumas rotinas e scripts para obtenção de arquivos com as bases foram criados usando a linguagem Python, de forma a se tornar mais ágil a obtenção das fontes de dados e verificações de lançamento de novas versões.

- Treinamento de uso para os pesquisadores do Núcleo: Algumas das ferramentas foram apresentadas para os pesquisadores que já costumam trabalhar com as ferramentas e aplicativos comuns, como o protocolo ODBC. Porém, como o sistema principal ainda está em fase de desenvolvimento, não houve oportunidade de promover treinamentos para todos no Núcleo.

Mudanças no plano de atividades:

Algumas das atividades que constavam no plano deste semestre foram repriorizadas, de modo a se tornarem atividades para a próxima etapa do desenvolvimento. Os motivos para esta decisão foram os seguintes:

- Excesso de horas despendidas nas atividades de limpeza, padronização e sincronização dos dados obtidos, devido ao grande volume de dados a ser processado.
- Uma pane no servidor Linux comprometeu o ambiente de desenvolvimento que estava sendo criado. Os dados não foram perdidos, porém o problema técnico ainda está pendente de ser sanado, com isso o trabalho se concentra só no servidor restante, o Windows, impossibilitando o desenvolvimento particular de algumas das ferramentas propostas inicialmente, como o *Portal Web* e o *componente de análises geográficas (SIG).*

Estes dois eventos, apesar de infortúitos, podem ser tomados como aprendizado de forma a otimizar o planejamento para as etapas futuras. Devido ao volume de trabalho, e visto que a atividade de agregar novas fontes ao banco de dados central é uma tarefa recorrente e para que a qualidade do banco seja mantida, uma sugestão seria a de agregar um novo bolsista técnico para colaboração nestas tarefas. O novo bolsita trabalharia seguindo as definições já estabelecidas para sincronia de novos dados com o banco de dados central, padronizado.

Neste último semestre de 2012 novos projetos ligados à tecnologia da informação foram implantados no NUPPs, por pesquisadores-associados. São eles a Corrupteca e o Blog Qualidade da Democracia. Estes dois projetos, principalmente o primeiro, consiste numa nova e rica fonte de dados possíveis de serem integradas também com o banco de dados central, para estarem de uma forma alternativa, disponíveis aos pesquisadores. Gera-se então uma nova demanda, o que justificaria também a agregação deste novo bolsista, que poderia acumular a função de desenvolver protocolos de integração entre os sistemas presentes no Núcleo.

Renovação de bolsa e solicitação de novo auxílio:

Pelos motivos apresentados na seção anterior, tanto para continuidade do projeto de desenvolvimento do sistema de informações do NUPPs, quanto para o trabalho contínuo de obtenção e agregação de novas fontes de dados para integrar o rol central do Núcleo, seguem as solicitações:

a) Renovação de bolsa de Capacitação Técnica nível TT4-a pelo período adicional de 12 meses, conforme plano de atividades em anexo;

b) Inclusão de uma bolsa adicional de Capacitação Técnica de nível TT3, também pelo período de 12 meses, conforme o plano de atividades em anexo.

Apreciação do desempenho do bolsista:

O relatório circunstanciado apresentado acima dá indicações claras do adequado cumprimento das tarefas atribuídas ao bolsista de Capacitação Técnica nível TTA-4. O processo de formação do banco geral de dados avançou bastante e começa a oferecer excelentes condições para a sua utilização pelos pesquisadores das diferentes áreas envolidas neste projeto de pesquisa. A avaliação do seu desempenho é, portanto, positiva e reconhece a relevância de sua contribuição para o desenvolvimento do projeto e, por isso, se solicita a extensão de sua bolsa de treinamento técnico.