5
Conclusão

Neste trabalho abordamos o tema da proveniência de dados em SGWC, com ênfase nas necessidades de experimentos in-silico de Bioinformática, como o uso de programas de linha de comando, arquivos locais e bases de dados grandes.

No capítulo 2 fizemos um levantamento de alguns desafios atuais na área de proveniência. Escolhemos alguns desses desafios para tratar neste trabalho, e apresentamos como algumas ferramentas de SGWC mais populares abordam esses temas.

No capítulo 3 foi proposto um projeto de proveniência priorizando alguns objetivos motivados pelos seguintes desafios: Modelagem de proveniência, gerência de dados consumidos e produzidos, descrição e gerência de atividades, reuso de dados e reprodutibilidade. Apresentamos as características resultantes da proposta em relação a cada desafio de interesse.

No capítulo 4 apresentamos a implementação do projeto como uma extensão do BioSide. Utilizamos dois workflows de Bioinformática como estudos de caso. O primeiro estudo de caso foi obtido diretamente dos exemplos disponibilizados no site do BioSide, e se mostrou interessante por poder ser completamente modelado usando o projeto de proveniência proposto. Já o segundo estudo de caso, MHOLline, é uma aplicação disponível na web, que suscitou alguns problemas interessantes e demonstra algumas limitações do modelo criado. Ambos usam ferramentas de linha de comando e bases de dados locais e grandes, característica interessante para este trabalho. Implementamos o modelo através de uma extensão ao sistema BioSide escrita em Java, permitindo a captura de dados de proveniência e o registro das mesmas em uma base de dados relacional (Postgres).

Nas próximas seções enumeramos as contribuições da dissertação e os possíveis trabalhos futuros.

5.1. Contribuições

As contribuições principais deste trabalho foram:

- Análise de alguns SGWC do ponto de vista de proveniência de dados. Existem muitos trabalhos que descrevem funcionalidades de proveniência dos sistemas. Procuramos analisar algumas características que são tratadas de maneira superficial nos trabalhos relacionados, como a descrição e gerência de atividades usadas para acessar programas stand-alone e o gerenciamento de dados.
- Proposta de um projeto de proveniência de dados que visa demonstrar como alguns desafios apresentados poderiam ser tratados nos SGWCs.
- Implementação do projeto proposto em um SGWC particular (BioSide) e teste do mesmo com 2 workflows de Bioinformática.
- Adaptação do MHOLline, com necessária descrição de atividades para execução do mesmo no BioSide.

Em relação aos desafios podemos listar as seguintes contribuições:

- Modelo de Proveniência: Os sistemas estudados produziram modelos pouco genéricos e muito dependentes da tecnologia utilizada no sistema. A proposta deste trabalho é representar informações de proveniência para workflows executados em SGWC de forma conceitual e independente de tecnologia. Nesse contexto consideramos importante utilizar o OPM para representar a proveniência retrospectiva.
- Reprodutibilidade: Classificação de dois níveis de reprodutibilidade e armazenamento de dados de proveniência que dão suporte limitado a esses níveis.
- Reuso de Dados: Proposta de uma forma de reutilização de artefatos gerados por uma execução específica de um workflow em outro workflow, mantendo a proveniência da origem do dado.
- Descrição e Gerência de Atividades: Os sistemas estudados não incluem no modelo de proveniência descrições sobre as atividades, mas apenas os passos do workflow. Consideramos importante a inclusão de descrições de atividades para a

proveniência e o reuso da definição de um workflow. Incluímos também descrições sobre programas de linha de comando invocados pelo workflow.

- Gerência de Dados Consumidos e produzidos: Foi elaborada uma modelagem que permite interligar execuções específicas aos seus dados consumidos e produzidos. A abordagem considerou o uso de bases de dados locais e arquivos grandes, sugerindo a participação do usuário para a decisão de armazenamento de uma cópia destes arquivos para reprodutibilidade estrita.

5.2. Trabalhos Futuros

- Melhoria no algoritmo de verificação de alteração de definição workflow, não realizando o versionamento se a definição foi alterada apenas em relação ao posicionamento dos elementos na interface.
- Elaborar um conjunto de consultas a proveniência, avaliando o esquema gerado em relação às diferentes consultas que podem ser realizadas.
- Avaliar o impacto da solução em relação à dificuldade de elaborar consultas à proveniência em uma base de dados relacional.
- Incluir registro da Origem do Dado como em [Guimaraes, 2009].
- Evolução para um componente de software independente.