Gerenciamento dos Dados

Ennio P Lopes

29 de abril de 2016

Este documento aborda tópicos de boas práticas para a coleta e gerenciamento de dados. Está fundamentado em discussões dos websites StatExchange, no tag do R em Stackoverflow e em experiências do grupo de estudos em análise de dados da EESC-USP.

Práticas

- Crie uma estrutura para o projeto:
  Tenha uma organização de pastas e arquivos para manter os dados, códigos, figuras, etc. nos lugares certos, de forma que o acesso fique intuitivo e fácil. Faça de uma maneira que possa evitar arquivos duplicados e utilize sistemas de controle de versão (git).
- Nunca modifique os dados originais:
  Mantenha os arquivos de dados brutos somente para leitura (se possível defina-os com o atributo read-only), copie e renomeie quando realizar transformações, limpezas, etc.
- Confira a consistência dos dados:
  Realize uma verificação se os dados importados estão em conformidade com a base original. A limpeza e transformações realizadas foram executadas conforme o esperado.
- Faça backup de tudo regularmente
  Esta é uma ação importante para a segurança dos arquivos.
- Mantenha um registro das suas ideias:
  (ou conte com um controlador de assuntos). Em parte redundante com o item anterior já que está disponível no git.
- Documente as etapas realizadas de data mining:
  Utilize editores e faça reports para registrar a origem e método de extração dos dados e as técnicas e etapas de manipulação. Faça o data mining reprodutível (ou pelo menos se esforce para conseguir o máximo de reprodutibilidade).
- Dicionário dos dados (codebook):
  É parte dos metadados no qual é utilizado para entendimento dos dados e da base de dados. Ele identifica elementos dos dados tais como nomes, definições e unidades de medidas e outras informações. Um dicionário de dados simples é uma coleção organizada dos nomes e definições dos elementos dos dados, organizados em uma tabela. Deve descrever toda a coleção dos dados de uma organização um parte da coleção ou uma única base de dados. Dicionário de dados mais avançados podem conter o esquema da base de dados com chaves de referência e diagrama entidade-relacionamento. Dicionários de dados fornecem uma janela para o conteúdo das bases de dados os quais possibilitam dar inicío ao processo de identificação do nível de similaridade entre bases.
  ISO 11179 é uma norma para padronizar e registrar elementos dos dados.

O desenvolvimento de conjuntos de elementos consistentes e formatações para documentação do conteúdo e estrutura da base de dados permite que os o sistema de informação seja mais acessível. Ferramentas do repositório dos metadados também devem estar disponíveis para manter o repositório do dicionário dos dados organizado como as fontes online, por exemplo: estrutura de tabelas; protocolo de coleta; elementos dos dados; termos e definições dos elementos dos dados.
Por fim o compartilhamento dos dados é melhorado quando usuários têm o acesso, tanto técnico quanto semântico dos dados, necessários para entender com profundidade as definições e pressupostos do sistema de informação.

Referências

http://stats.stackexchange.com/
http://stackoverflow.com/questions/tagged/r MEDEIROS, D. General Use Data Delivery Codebook/Data Dictionary SONG, W. et al. An efficient method to create a large and comprehensive codebook