Bancos de Dados e Observatório Virtual

Relatório Resumido
Sub-comissão da CEA

17 de março de 2010

Albert Bruch (Relator), Claudio Bastos Pereira, Haroldo Campos Velho, Luiz Nicolaci da Costa, Paula R.T. Coelho, Reinaldo de Carvalho

A astronomia tornou-se, nas últimas décadas uma ciência que gere um imenso fluxo de dados que, na era digital, são disponíveis em sistemas de banco de dados ao redor da Terra e que já somam petabytes. A taxa de geração de dados continua crescendo exponencialmente, dobrando a cada ~18 meses. Além da quantidade, observa-se um aumento da complexidade dos dados. Com o horizonte de novos grandes projetos da astronomia observacional, fica óbvio que esse desenvolvimento se manterá no futuro previsível.

O eficiente uso desses dados, maximizando a extração do seu enorme conteúdo científico e aproveitando as possibilidades tecnológicas existentes hoje para ampliar o escopo espacial, espectral e até temporal de pesquisas astronômicas através de estudos sinérgicos utilizando múltiplos acervos com dados obtidos com diversos observatórios e instrumentos (terrestres e no espaço), em faixas espectrais de raios gama até ondas de rádio, cobrindo todo o céu e tomadas em épocas distintas, exige novas formas para realizar pesquisas astronômicas. O potencial do uso dos numerosos bancos de dados astronômicos complementa e ultrapassa a forma tradicional da pesquisa astronômica, a saber: o estudo individual de conjuntos relativamente pequenos de dados. Implica em uma nova forma de se fazer astronomia, que apresenta nada menos do que um novo paradigma para a astronomia mundial.

Esse conceito diferente e inovador levou a comunidade astronômica a considerar o conjunto de acervos de dados como um enorme“Observatório Virtual”(OV) e levou a esforços para criar tecnologias e ferramentas com intuito de permitir e facilitar o uso integrado dos mesmos para pesquisa, aplicando o novo paradigma mencionado acima; esforços esses coordenados mundialmente pela Internacional Virtual Observatory Alliance – IVOA.

No Brasil, o uso do Observatório Virtual e o aproveitamento da enorme quantidade de dados disponíveis e já oferecida de forma organizada em numerosos servidores no mundo inteiro ainda está incipiente. Os motivos incluem a ausência de uma infra-estrutura adequada e o enfoque dos pesquisadores em métodos de pesquisa mais tradicionais, que não tiram proveito do potencial oferecido pelos grandes bancos de dados existentes e da estrutura de OV criada nos últimos anos. Existem alguns grupos no país focados na criação de uma infra-estrutura de banco de dados, desenvolvimento de ferramentas para o uso de dados arquivados no contexto de OV, e geração de uma cultura para o uso do potencial científico oferecido pelo OV na comunidade astronômica nacional. Mais notadamente mencionamos aqui:

- O projeto Astrosoft, radicado no Observatório Nacional – ON R e com participação do Centro Brasileiro de Pesquisas Físicas – CBPF que visa a desenvolver uma plataforma para o armazenagem e o processamento de grandes quantidades de dados e sua disponibilização para usuários.

- O BraVO (Brazilian Virtual Observatory), criado no contexto do Instituto Nacional de C&T de Astrofísica, que reúne, num primeiro instante, grupos de trabalho do Instituto Nacional de Pesquisas Espaciais – INPE, do Laboratório Nacional de Astrofísica – LNA, da Universidade de São Paulo – USP e da Universidade Federal de Santa Catarina – UFSC. O BraVO tem como intuito a criação de uma infra-estrutura compartilhada para o processamento e a mineração de dados arquivados, o desenvolvimento de ferramentas para o VO e o treinamento de pesquisadores para o uso eficiente do potencial de OV. O BraVO colabora, para esses fins, com a IVOA.

Identificamos três áreas principais onde se faz necessário atuar para tornar eficaz e eficiente o aproveitamento do potencial científico inerente aos grandes bancos de dados existentes e para preparar a comunidade astronômica para extrair o conteúdo científico do fluxo de dados provindos de novas máquinas geradoras de dados (em implementação ou projetados, tais como, o SDSS-III, DES, PanSTARRS, LSST, VISTA e outros, alguns deles já com participação brasileira direta). Em todas as áreas, os investimentos financeiros necessários para realizar uma melhoria decisiva da infra-estrutura de VO no Brasil são relativamente modestos em comparação com outros grandes projetos da astronomia brasileira, ou irão beneficiar também diretamente outras áreas estratégicas da ciência no país (infra-estrutura da internet). São essas as áreas identificadas:

1. Recursos Humanos

   A geração de imensas quantidades de dados e a subsequente necessidade de armazená-los e o desenvolvimento e aplicação de ferramentas para a extração de informação científica destes dados tem origem em uma demanda dos astrônomos. Porém, enquanto a interpretação dos dados é atribuição dos profissionais em astronomia, os passos anteriores (armazenamento em banco de dados, criação de ferramentas de análise, técnicas de mineração de dados, processamento distribuídos e aplicativos executados via interner) requerem conhecimentos específicos principalmente de áreas da informática, na qual o astrônomo geralmente não tem uma capacitação específica. Existe uma camada de interface na definição e criação de ferramentas de alto nível em que astrônomos deverão definir as características e a funcionalidade das mesmas e poderão colaborar com profissionais da computação no seu desenvolvimento e implementação. Entretanto, a estruturação e a realização de grandes bancos de dados, a definição de protocolos e a implementação de uma estrutura para o uso compartilhado dos arquivos é tarefa inerente da informática, enquanto que o uso das ferramentas e interpretação dos resultados cabe aos astrônomos. Isso gera a necessidade de atuar, no que se refere aos recursos humanos, em duas frentes:

   a) Enquanto os astrônomos em geral já estão conscientes do potencial científico da mineração de dados, o uso em maior escala de bancos de dados como ferramenta para pesquisa apresenta um novo paradigma na astronomia. Não apenas aceitar isso como fato, mas internalizar um novo paradigma para torná-lo uma forma normal no pensamento de um pesquisador é um processo lento. Geralmente demora uma geração para que isso aconteça: para que o novo paradigma seja absorvido plenamente pela cultura científica ele deverá ser parte da formação do profissional desde o início. Portanto, há necessidade de não apenas capacitar os pesquisadores em astronomia no uso de grandes bancos de dados e das ferramentas do OV, habituando os mesmos a uma nova forma para fazer pesquisa, mas também de incorporar o novo paradigma na profissionalização dos estudantes de astronomia.

b) Criar uma estrutura seguindo plenamente as ideias do Observatório Virtual como uma grande rede interligando os mais diversos arquivos de dados astronômicos certamente é um enorme desafio para a tecnologia da informação e, portanto, deverá atrair os profissionais da área. Entretanto, é preciso informar e motivar a comunidade de computação sobre este novo desafio. Cabe aos interessados no resultado final, os astrônomos, chamar a atenção dos tecnologistas em informática para colaborar nesse grande empreendimento. Isso exige de um lado, uma atitude proativa dos astrônomos para se aproximar e interagir fortemente com profissionais da área da informática e, de outro lado, criar mecanismos para atrair profissionais, que vão além de simplesmente apontar o desafio interessante. É preciso ter em conta que existem outros desafios competindo pela atenção deste tipo de profissional, sem mencionar que a iniciativa privada oferece outros atrativos (p.ex., financeiros) principalmente para os mais capacitados na tecnologia da informação.

2. Envolvimento de instituições externas da área da astronomia

A necessidade de uma forte interação entre a astronomia e a informática e disciplinas afins no desenvolvimento de uma estrutura de OV, torna evidente que as finalidades poderão ser atingidas apenas através de um forte envolvimento de instituições externas da área da astronomia. Já existem colaborações nesse sentido entre o Projeto Astrosoft e o Laboratório Nacional de Computação Científica – LNCC e por meio da participação do Laboratório Associado de Computação e Matemática Aplicada – LAC/INPE no BraVO. Há também contatos iniciais entre o BraVO e o LNCC. Certamente ainda existe um grande potencial de crescimento para essas colaborações interdisciplinares.

Apresentada a necessidade de ampliar capacidades de rede de internet (veja próximo item), um forte empenho da Rede Nacional de Ensino e Pesquisa (RNP) é indispensável. Outras instituições no âmbito da estrutura do MCT com competência para contribuir para o desenvolvimento de tecnologia para bancos de dados e OV incluem o Centro da Tecnologia da Informática Renato Archer (ICT). Além disso, existem instituições, laboratórios e grupos de pesquisa em universidades brasileiras, algumas delas – mas certamente não todas – já identificadas pelos astrônomos interessados no assunto, cujo potencial para colaborar merece ser investigado.

Apesar do fato de que colaborações entre a área acadêmica – principalmente atuando em pesquisa básica como a astronomia – e a indústria privada não tem muita tradição no Brasil, não deverá ser descartada a colaboração de empresas do ramo da informática ou diretamente com os astrônomos ou com as instituições acima mencionadas no desenvolvimento da infra-estrutura para o OV. Exemplos no exterior mostram o caminho (p.ex., a participação do Google Inc. no Large Synoptic Survey Telescope). Um incentivo para as empresas colaborarem no presente assunto poderá ser o efeito de sinergia, considerando que as tecnologias desenvolvidas no contexto do OV tem aplicações nas mais diversas áreas, muitas delas de interesse comercial. Além do atual contexto, outros incentivos para as empresas colaborarem mais fortemente com a área acadêmica em geral merecem ser criados (p.ex., no âmbito do SIBRATEC).

3. Infra-estrutura

Para o bom aproveitamento das oportunidades científicas oferecidas pela disponibilidade de grandes acervos de dados, precisa-se de uma infra-estrutura que ou ainda não existe no Brasil

ou ainda carece da configuração e organização necessárias para seu uso eficiente e eficaz. Podemos distinguir diversas áreas onde é preciso investir em infra-estrutura:

a) Capacidade de rede

Enquanto a ideia do OV parte da premissa que a análise dos dados pesquisados acontece no local onde eles se encontram e apenas os resultados serão transferidos para o usuário remoto (evitando a necessidade de transferência de grandes quantidades de informação), a realidade do OV ainda não chegou a esse ponto. Por enquanto, vigora ainda um padrão de uso descentralizado até de dados idênticos, com espelhos de acervos em vários locais, exigindo o transporte de muitos dados pela internet e, portanto, altas capacidades de transmissão de dados pela rede. Mesmo quando se chega a um modelo no qual múltiplos espelhos de acervos se tornam desnecessários, o transporte de dados do local onde eles são gerados até o local de armazenamento (muitas vezes distinto) requer boas conexões de rede. Finalmente, processamento de dados em rede (veja abaixo), usando processadores distribuídos em locais geograficamente diferentes, somente são eficientes mediante um bom e rápido intercâmbio de informações entre eles.

É atribuição da RNP fornecer e gerenciar as redes da internet utilizadas pelas instituições de pesquisa e ensino do Brasil. A conectividade depende do local, sendo maior em algumas Capitais e cidades com grandes instituições e menor no interior dos Estados. Há um anel de conexão entre as cidades de Brasília, São Paulo, Rio de Janeiro e Belo Horizonte de boa velocidade. Conectando-se a esta rede existem redes de menor poder de tráfego de dados para o Nordeste e Sul do País. É preciso colaborar com a RNP para garantir celeridade na implementação de redes rápidas, principalmente entre aqueles centros de pesquisa onde hardware computacional de capacidade de armazenamento de dados para uso compartilhado da comunidade astronômica já existe ou será instalado (ver abaixo). Evidentemente isso iriá beneficiar também outras áreas da pesquisa, fator que deverá facilitar a criação das capacidades necessárias.

b) Hardware computacional

Um levantamento preliminar feito pelo BraVO mostra que as capacidades em hardware computacional disponíveis nas instituições astronômicas estão longe de satisfazer as demandas impostas pelo novo paradigma de OV, para tratamento e análise de dados astronômicos. Investimentos significativos são necessários para que os pesquisadores brasileiros possam tirar proveito das oportunidades oriundas dos grandes projetos astronômicos do futuro (próximo!). Precisa-se ainda de um estudo cuidadoso das reais necessidades para evitar investimentos errados. Isso inclui a eventual elaboração de modelos para uso compartilhado de hardware, computação em grade (com os processadores centrados em um único lugar ou distribuídos e interconectados através da internet) e aplicação de novas tecnologias e arquiteturas computacionais. Ainda deverá se investigar se existem hardware e estruturas computacionais em instituições externas da área de astronomia, disponíveis e adequadas para o uso no contexto em pauta (p.ex., o SINAPAD).

Para facilitar a implementação de novo hardware computacional os astrônomos deverão ter acesso a recursos das agências de fomento reservados para a área da informática.

c) Software

Software nos mais diversos níveis é essencial para o aproveitamento de grandes acervos de dados no âmbito do VO. Isso inclui:

- Software de alto nível (ferramentas) para a aplicação de tarefas específicas nos dados (análise de imagens, de espectros, etc.). Desenvolvimento nessa área está sendo realizado tanto no Astrosoft quanto no BraVO por astrônomos em colaboração com profissionais da informática.

- Desenvolvimento de algoritmos eficientes para computação em grade (caso os estudos acima mencionados mostrem a importância disso). Criar plataforma de software desse gênero já transcende a competência de astrônomos e necessita da colaboração de instituições externas com capacitação na área ou a participação direta em projetos (via contratação, bolsas, etc) de especialistas na área.

- Software de baixo nível, p.ex., para viabilizar a interoperabilidade de bancos de dados. Enquanto isso é um assunto tratado por diversos grupos da IVOA, não tem atividades no Brasil nesse sentido.

Como complemento ao desenvolvimento de novos pacotes de software para aplicações específicas, deverá ser contemplado o uso de grandes ambientes de software geral, tais como o sistema Astrowise, com ampla aplicação e utilização no gerenciamento de bancos de dados e no contexto do OV.

d) Capacidade de armazenamento de dados

Como já foi dito, a geração de dados na astronomia cresce de forma exponencial. Para disponibilizar a enorme quantidade de informação à comunidade científica, seja de forma original ou via espelhos distribuídos, precisa-se obviamente de capacidades de armazenagem que crescem com a mesma taxa. Considerando a impossibilidade do Brasil abrigar ou espelhar os dados de todos os grande geradores de dados do futuro, os investimentos necessários dependerão dos projetos nos quais a astronomia brasileira terá um envolvimento direto. Portanto, é preciso um levantamento cuidadoso das reais demandas em armazenamento de dados.

A infra-estrutura acima mencionada será, em grande parte, para o uso compartilhado da comunidade astronômica. Portanto, precisa-se de uma política que garanta a todos os interessados o acesso e deverá também incluir uma avaliação do uso e um planejamento para sua implementação, atualização e planos de ampliação, ou seja, um modelo de gerenciamento da infra-estrutura compartilhada. A melhor forma para efetuar isso ainda precisa ser discutida.