Um Modelo de Estrutura Organizacional em Plataformas de E-Science

Glauber J. Vaz, Poliana F. Giachetto, Tércia Z. Torres, Silvia M. F. S. Massruhá

Embrapa Informática Agropecuária
Caixa Postal 6041 – 13083-886 – Campinas – SP – Brasil

{glauber, poliana, tercia, silvia}@cnptia.embrapa.br

Abstract. *More attention should be given to the organizational structure in e-Science platforms. The model of organization adopted by Multi-User Bioinformatics Laboratory of Embrapa have shown to be efficient in offering a collaborative research environment and infrastructure to bioinformatics groups from Embrapa and partner institutes. Here, we present our model, that can be used as a reference for building an e-Science platform in other institutions.*

Resumo. *Uma maior atenção deveria ser voltada à infraestrutura organizacional nas plataformas de e-Science. O modelo de organização adotado pelo Laboratório Multiusuário de Bioinformática da Embrapa tem se mostrado eficaz no oferecimento de um ambiente colaborativo e de uma infraestrutura científica para grupos de pesquisa que trabalham com Bioinformática na Embrapa e instituições parceiras. Aqui nós apresentamos esse modelo, que pode ser utilizado como referência por outras instituições na construção de uma plataforma de e-Science.*

1. Introdução

O conceito de plataforma de e-Science para o avanço da pesquisa, segundo relatório do *Research Councils UK*[1] (2009), considera não somente os sistemas que atendem às definições tradicionais da computação (software), mas também outras formas importantes de infraestrutura. As estruturas organizacionais (grupos formais e informais que oferecem serviços de e-Science), assim como o capital humano (conhecimento e experiência) e os recursos de dados e informações (sistemas que suportam o crescente volume de dados gerados pela pesquisa), exercem papel de grande relevância em uma plataforma de e-Science.

Este trabalho trata da estrutura organizacional em plataformas de e-Science. Descreve a organização do Laboratório Multiusuário de Bioinformática (LMB) da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e mostra que o modelo de organização adotado pelo LMB favorece a construção de soluções eficazes para o

1 *Research Councils UK* (2011) – RCUK - é uma parceria estratégica entre os Conselhos de Pesquisa do Reino Unido criada em 2002 para permitir que eles trabalhem colaborativamente de maneira mais eficaz e gerem maior impacto de suas atividades de pesquisa, treinamento e inovação, contribuindo assim, na realização dos objetivos governamentais para a Ciência e a inovação. O termo e-Science foi cunhado por John Taylor, que foi Diretor-Geral dos Conselhos de Pesquisa do Reino Unido.

enfrentamento dos desafios impostos pela Ciência a instituições de pesquisa. Este modelo, portanto, pode ser utilizado na constituição de outras plataformas de e-Science.

O LMB tem por missão viabilizar soluções de Bioinformática para projetos de pesquisa, desenvolvimento e inovação na Embrapa em um ambiente colaborativo. Sediado na Embrapa Informática Agropecuária, conta com a atuação de profissionais de diversas áreas do conhecimento, distribuídos nas dezenas de unidades da empresa, e ainda com a colaboração de profissionais de outras instituições de pesquisa. Seus principais objetivos são: (i) contribuir para o avanço na fronteira do conhecimento e incorporar novas tecnologias em Bioinformática; (ii) viabilizar soluções eficientes para demandas em Bioinformática; (iii) disponibilizar acesso à infraestrutura computacional de alto desempenho; e (iv) prover a Embrapa de competências em Bioinformática [Embrapa 2011b]. O LMB, portanto, constitui uma plataforma de e-Science na Embrapa para a área de Bioinformática. A fim de cumprir seu papel, é fundamental que apresente uma infraestrutura organizacional que facilite o atendimento às principais demandas dos cientistas.

Este trabalho está estruturado em quatro seções contando com esta introdução. A segunda seção contextualiza, a partir de Vaz (2011), o conceito de e-Science, na Embrapa e no Brasil, e mostra que os laboratórios multiusuários, conforme concebidos pela Embrapa, são estratégicos nestes dois cenários, tanto na empresa quanto no país. Na terceira seção, apresenta-se o modelo de gestão do LMB e alguns dos seus principais desafios no que se refere mais especificamente à área de Computação nesta fase inicial de implantação. As conclusões são apresentadas na última seção.

2. E-Science na Embrapa e no Brasil

Propostas relacionadas a e-Science estão totalmente alinhadas às diretrizes estratégicas não só da Embrapa como do país. O chamado Livro Azul sintetiza as principais contribuições da 4ª Conferência Nacional de Ciência, Tecnologia e Inovação (CNCTI) para o Desenvolvimento Sustentável (2010b), convocada por decreto presidencial para discutir uma política de Estado para ciência, tecnologia e inovação com vistas ao desenvolvimento sustentável. No Livro Azul, afirma-se que nenhum país que aspire a ser moderno e desenvolvido pode abrir mão de investir seriamente na área de Tecnologias de Informação e Comunicação (TICs). Além disso, e-Science é explicitamente citada como foco de investimentos no documento que consolida as recomendações da 4ª CNCTI (2010a). Portanto, e-Science está em evidência nas discussões que envolvem o desenvolvimento sustentável no Brasil.

As estratégias da Embrapa também estão em consonância com os princípios de e-Science e com as diretrizes estipuladas para a ciência, tecnologia e inovação no país. Para a ampliação e modernização da infraestrutura de pesquisa, por exemplo, foram enumeradas as seguintes recomendações na 4ª CNCTI [2010a]: (i) acréscimo de investimentos em infraestrutura; (ii) criação de novas instalações de uso multi-institucional, especialmente em áreas estratégicas, como a biotecnologia; (iii) fortalecimento do papel das unidades de pesquisa ligadas aos ministérios que forneçam a infraestrutura adequada, promovam ações estruturantes e sirvam de âncoras para grandes projetos científicos e tecnológicos de interesse da sociedade brasileira. A Embrapa, vinculada ao Ministério da Agricultura, Pecuária e Abastecimento, já lidera

grandes projetos científicos nacionais, mas ainda tem muito a contribuir. Pode, por exemplo, ampliar suas redes de pesquisa, melhorar sua infraestrutura científica e disponibilizá-la a parceiros. Em seu Plano Diretor [Embrapa 2008], destacam-se as seguintes estratégias: (i) assegurar o uso compartilhado de equipamentos, laboratórios, informações e campos experimentais entre técnicos, Unidades Descentralizadas, pesquisadores e parceiros para assegurar a manutenção, a atualização e a melhor utilização da infraestrutura laboratorial, de tecnologia da informação e dos campos experimentais; (ii) estimular o compartilhamento de infraestrutura em laboratórios multiusuários de excelência; (iii) articular o ambiente de cooperação em rede; e (iv) articular redes cooperativas, produtivas e sociais, com base em modelos de gestão ágeis e flexíveis para a construção de plataformas tecnológicas. Além disso, o Plano Diretor da Embrapa sugere a intensificação de PD&I em temas estratégicos para o Brasil e cita, entre eles, a bioinformática e a tecnologia da informação.

A criação de laboratórios multiusuários é uma resposta da empresa a esses desafios relacionados à infraestrutura científica. Para a Embrapa (2011a), os laboratórios multiusuários são aqueles utilizados para realizar testes e análises de alta complexidade científica, envolvendo equipe técnica multidisciplinar e equipamentos altamente especializados. Estes laboratórios devem funcionar como plataformas compartilhadas de pesquisa, disponíveis para utilização por múltiplas unidades da empresa, além de instituições parceiras. Desta maneira, integram-se amplas redes de pesquisadores em torno de infraestruturas modernas, geridas com agilidade e dinamismo para incorporar o que existe de melhor, seja em instrumentação e equipamentos laboratoriais, seja em competências para sua operação.

O LMB é a primeira experiência da Embrapa na implantação desses laboratórios multiusuários. Entre os fatores que determinaram essa escolha, está o fato de que a Bioinformática aplicada à Agricultura tem se destacado como uma área estratégica, o que pode ser verificado por algumas das recomendações feitas na 4ª CNCTI relacionadas à agricultura e à bioinformática: (i) Fortalecimento da infraestrutura de C,T&I para o setor agrícola; (ii) Apoio ao desenvolvimento de atividades de P&D em nanotecnologia, biotecnologia e bioinformática para agricultura; (iii) Desenvolvimento de pesquisas de melhoramento genético.

Com a implantação do LMB, projetos científicos que demandam computação de alto desempenho tornaram-se viáveis na Embrapa. O projeto “Rede de sequenciamento de genomas para desenvolvimento de tecnologias inovadoras para a pecuária brasileira”, por exemplo, envolve a realização da montagem *de novo* do *Bos indicus* (Nelore). Seu genoma é ainda maior do que o do Panda gigante, montado recentemente [Li et. al. 2010] com a utilização de um supercomputador de 32 núcleos e 512 GB de memória RAM. A infraestrutura do LMB, descrita adiante em maiores detalhes, também oferece recursos para a realização desta tarefa. Este projeto, que envolve a montagem do genoma bovino, é apenas um exemplo, pois há muitos outros que envolvem espécies vegetais e animais de interesse para a Embrapa: suínos, aves, dendê, eucalipto, arroz, café, soja, milho e sorgo, dentre outras. Portanto, o LMB tem condições, atualmente, de viabilizar projetos científicos na área de bioinformática que demandam recursos computacionais de alto desempenho. Isso contribui ainda mais para que a Embrapa se posicione na fronteira do conhecimento na área de bioinformática aplicada à

agropecuária.

A experiência com o LMB deve ser estendida a outros campos em que a Embrapa atua. A partir deste caso, será mais fácil identificar elementos comuns à implantação de novos laboratórios multiusuários. Assim, será possível atender melhor às demandas por infraestrutura científica da empresa toda e de parceiros. Outras instituições de pesquisa, ao construir suas próprias plataformas de e-Science, podem utilizar o LMB como uma referência.

Um dos primeiros passos para a constituição do LMB foi a definição de sua estrutura organizacional.

3. Estrutura organizacional do LMB

O LMB é gerenciado por um responsável técnico, que o coordena com o auxílio de um Comitê Assessor, de caráter consultivo. Dado que o LMB envolve a participação de vários Centros de Pesquisa da empresa, além de instituições externas, tanto o coordenador quanto o Comitê Assessor são designados pelo próprio Diretor de Pesquisa e Desenvolvimento da Embrapa, a quem compete a definição e divulgação das condições de implantação, acompanhamento e avaliação da operacionalização do funcionamento do LMB, de modo a viabilizar uma gestão ágil e flexível. Seu coordenador também responde à chefia da Embrapa Informática Agropecuária.

Três macroprocessos principais foram estabelecidos para o LMB: Gestão Global, Gestão de Tecnologia da Informação (TI) e Gestão do Negócio. A Figura 1 ilustra sua estrutura organizacional.

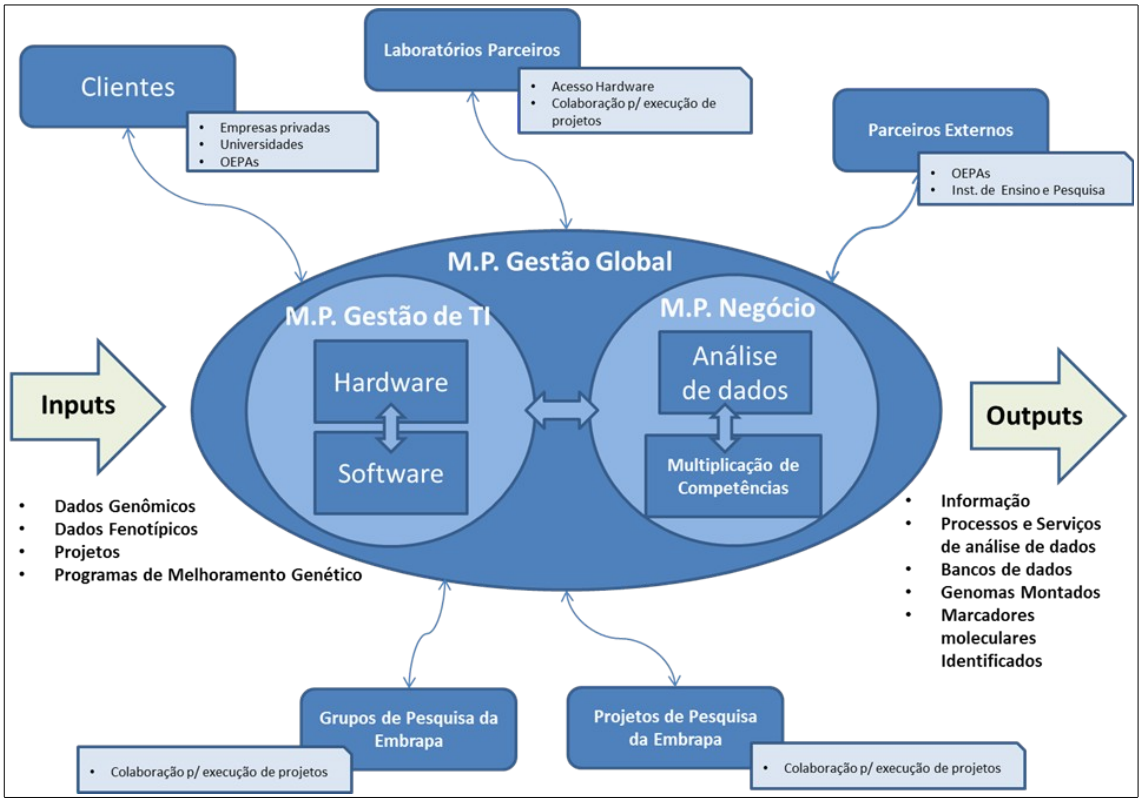

Figura 1. Estrutura organizacional do LMB.

O macroprocesso de Gestão Global envolve os processos relacionados à interface do LMB com os agentes internos e externos, como empresas privadas, universidades e instituições de pesquisa. Além disso, garante a sinergia entre os outros dois macroprocessos.

No macroprocesso de TI, há dois processos principais, Gestão de Hardware e Gestão de Software, que garantem as condições necessárias para o bom funcionamento do LMB em termos de infraestrutura básica de TI.

Os processos de Análise de Dados e de Multiplicação de Competências são o núcleo do macroprocesso Gestão do Negócio. É função do LMB dar suporte aos projetos de pesquisa da Embrapa e parceiros nos diferentes contextos da Bioinformática, que envolvem desde análises de dados de *microarray* até a montagem de genomas complexos de animais e plantas. O LMB também procura transferir aos parceiros os conhecimentos necessários para que possam realizar suas próprias análises.

Para uma melhor compreensão desta organização, são apresentados a seguir alguns dos principais desafios que têm sido foco no desenvolvimento do LMB, associados às partes apresentadas na Figura 1.

Em primeiro lugar, consideram-se os ambientes virtuais de pesquisa (AVP). Um AVP pode ser visto como um *framework* em que ferramentas, serviços e recursos podem ser conectados [Fraser 2005], tais como: *wikis*, sites informativos, mecanismos de busca, software, serviços de armazenamento e de computação, canais de comunicação, fóruns de discussão, ferramentas de *workflow*, entre muitos outros. O LMB possui um AVP (www.lmb.cnptia.embrapa.br) em estágio inicial, cujo principal papel é potencializar a comunicação entre seus membros, clientes e parceiros, além de facilitar e estimular a pesquisa colaborativa. Na Figura 1, o AVP está presente em todas as setas, que representam a comunicação entre os atores envolvidos, e tem um papel de maior destaque na porção inferior da figura, que envolve a pesquisa colaborativa.

Os AVPs estão relacionados ao que Senge (2010) chama de infraestrutura que aprende, cuja construção deve ser guiada pela estratégia de se focar nas maneiras como a tecnologia pode auxiliar as pessoas de toda a empresa a encontrar umas às outras e a se ajudarem. A construção e a ampliação das redes de conhecimento além das fronteiras organizacionais também devem ser viabilizadas por uma infraestrutura que aprende.

A ampliação dessas redes é uma busca constante do LMB, uma vez que os desafios enfrentados atualmente são muito complexos e exigem composição de equipes transdisciplinares e multi-institucionais. Então, uma plataforma de e-Science deve evoluir de maneira a incluir muitos parceiros, nacionais e internacionais. Além de disponibilizar sua infraestrutura computacional para vários parceiros, o LMB elabora e executa projetos de pesquisa com outras instituições e agrega colaboradores das diversas unidades da Embrapa e de universidades, contando com docentes e alunos de todos os níveis, desde a graduação até o pós-doutorado.

O processo de Análise de Dados está no núcleo da Gestão do Negócio e atende às principais demandas ao LMB. Como os dados vão desde imagens de *microarrays* de expressão até arquivos texto contendo centenas de milhões de sequências geradas pelas novas tecnologias de sequenciamento, as análises são de diferentes naturezas e

demandam conhecimentos de muitas disciplinas. Matemática, estatística, física, computação e biologia são apenas algumas das mais importantes.

Um dos principais desafios está na organização e no armazenamento do grande volume de dados em questão. O próprio resultado das análises também é de grande volume, geralmente da ordem de dezenas de gigabytes. Além disso, algumas análises como, por exemplo, montagem *de novo* de genomas de organismos complexos, exigem memória principal medida em terabytes. Estes aspectos já dão a dimensão da complexidade das análises realizadas pelo LMB. Outro desafio importantíssimo é entregar esses resultados de tal forma que um não especialista em computação possa manuseá-los e interpretá-los da forma mais simples possível.

O LMB busca na sua própria equipe ou na rede de parceiros as competências necessárias para cada tipo de análise. Sempre buscando o estado-da-arte, diferentes estratégias são avaliadas até se chegar a *workflows* que sejam cientificamente robustos e computacionalmente integrados. Cabe ressaltar que alguns *workflows* estão sendo desenvolvidos por parceiros externos e são disponibilizados pelo LMB.

Verifica-se, então, uma estreita relação entre os processos de Análise de Dados e de Multiplicação de Competências, que se refere ao aprendizado das equipes do LMB e de parceiros. Cursos, eventos, documentações e o cuidado com a manutenção de parcerias são práticas comuns para se manter as competências no laboratório.

O macroprocesso de Gestão do Negócio depende completamente da Gestão de TI, que garante a infraestrutura básica para o funcionamento do LMB. Na Gestão de Hardware, os primeiros passos essenciais foram dados no sentido de disponibilizar poder de processamento e capacidade de armazenamento aos bioinformatas da Embrapa e parceiros, além de uma infraestrutura de comunicação. Devido à geração de uma quantidade crescente de dados e ao fato dos dados serem coletados em diferentes pontos geográficos, no futuro, provavelmente, será melhor enviar computação aos dados em vez de levar os dados à computação, como ocorre hoje no LMB. Esta é uma das regras informais que Gray formulou como princípios para projetar os sistemas do futuro baseados em grande volume de dados [Szalay and Blakeley, 2009]. Apesar de não ser viável no presente momento, esta abordagem deve ser levada em consideração em decisões que tenham efeito de longo prazo.

O parque computacional do LMB conta atualmente com dois servidores IBM System x3850 X5 (512 GB de memória RAM expansível para 6 TB, oito processadores Xeon 6-core E7540 com 2 GHz e cache de 18 MB, oito HDs SAS de 300 GB organizados em RAID1 e RAID5, disponibilizando 570 GB e 820 GB, respectivamente), um servidor HP ProLiant DL785 G6 (256 GB de memória RAM expansível para 512 GB, quatro processadores Six-core AMD Opteron™ 8431 com 2,4 GHz, quatro HDs SAS de 500 GB organizados em RAID5, disponibilizando 1,5 TB), um *storage* IBM DS3512 (60 Hds Sata de 2 TBs organizados em RAID5, disponibilizando 101 TBs) e um sistema de *backup*, com um servidor SunFire X4440 (32 GB de memória RAM, quatro processadores Quad-core AMD Opteron™, 400 GB de espaço para armazenamento em disco) e uma biblioteca de fitas IBM TS3200 (capacidade para 48 fitas LTO5, possibilidade de armazenamento de até 72 TBs *online* e software gerenciador de *backups* ARC Server Back).

Toda essa capacidade de armazenamento também influencia o desenvolvimento de software. Devido à necessidade de processamento de dados em escala de ordem superior a terabytes, novos algoritmos mais eficientes também precisam ser desenvolvidos [Bell et al. 2005]. A Gestão de Software cuida do desenvolvimento de novas ferramentas e seleção daquelas que já atendem às necessidades atuais.

4. Conclusões

As diretrizes expressas nos documentos estratégicos da Embrapa e do país rumo ao desenvolvimento sustentável estão alinhadas aos princípios da e-Science, que estabelece o compromisso de se atender da melhor maneira possível às necessidades dos pesquisadores, principalmente, no que se refere a recursos computacionais. Os laboratórios multiusuários da Embrapa, cujo pioneiro é o LMB, procuram atender a estas demandas da empresa e de parceiros. Em essência, o LMB constitui uma plataforma de e-Science para a área de Bioinformática.

Uma infraestrutura organizacional eficaz é fundamental em plataformas de e-Science para possibilitar o desenvolvimento das atividades de uma maneira eficiente e garantir seu funcionamento e desenvolvimento a longo prazo. A estrutura organizacional utilizada no LMB tem se mostrado eficaz no oferecimento de um ambiente colaborativo e de uma infraestrutura científica para grupos de pesquisa que trabalham com Bioinformática na Embrapa e instituições parceiras. Ela pode ser utilizada como modelo para se determinar a estrutura organizacional de novas plataformas de e-Science, não só na Embrapa como em outras instituições.

Agradecimentos

Agradecemos às inúmeras contribuições da equipe do LMB e, especialmente, às do Dr. Michel Eduardo Beleza Yamagishi.

Referências

Bell, G., Gray, J. and Szalay, A. (2005) “Petascale Computational Systems: balanced cyberInfrastructure in a Data-Centric World”, http://research.microsoft.com/en-us/um/people/gray/, 08 fev. 2011.

Conferência Nacional de Ciência, Tecnologia e Inovação para o Desenvolvimento Sustentável, 4 (2010a) “Consolidação das recomendações”. Brasília, DF: Ministério da Ciência e Tecnologia: Centro de Gestão e Estudos Estratégicos, http://www.cgee.org.br/publicacoes/livroazul.php, 31 jan. 2011.

Conferência Nacional de Ciência, Tecnologia e Inovação para o Desenvolvimento Sustentável, 4 (2010b) “Livro azul”. Brasília, DF: Ministério da Ciência e Tecnologia: Centro de Gestão e Estudos Estratégicos, http://www.cgee.org.br/atividades/redirect.php?idProduto=6820, 31 jan. 2011.

Embrapa (2011a) “Laboratórios Multiusuários da Embrapa” Síntese preparada pela Diretoria Executiva de Pesquisa e Desenvolvimento.

Embrapa (2011b) “Plano de implantação e operação do Laboratório Multiusuário de Bioinformática da Embrapa”. Boletim de Comunicações Administrativas, Brasília,

DF, Ano 37, n. 50, p. 10-12, out. 2011b. Resolução Normativa n. 16, set. 2011.

Embrapa (2008). Secretaria de Gestão e Estratégia. “V Plano Diretor da Embrapa: 2008-2011-2023”. Brasília, DF. 44 p.

Fraser, M. (2005) “Virtual research environments: overview and activity”. Ariadne, n. 44, Jul. 2005, http://www.ariadne.ac.uk/issue44/fraser, 31 jan. 2011.

Li, R. et. al. (2010) “The sequence and de novo assembly of the giant panda genome”. Nature, n. 463, p. 311-317.

Research Councils UK (2009) “RCUK Review of e-Science 2009”. Report of the International Panel for the 2009 Review of the UK Research Councils e-Science Programme, http://www.epsrc.ac.uk/pubs/reports/Pages/internationalreviews.aspx, 08 fev. 2011.

Senge, P. M. (2010) “A quinta disciplina: arte e prática da organização que aprende”. 26. ed. rev. ampl. Rio de Janeiro: BestSeller. 530 p.

Szalay, A. S. and Blakeley, J. A. (2009) “Gray’s Laws: Database-centric Computing in Science”. In: Hey, T., Tansley, S. and Tolle, K. (Ed.). The Fourth Paradigm: data-intensive scientific discovery. Redmond, WA, USA: Microsoft Research, p. 5-11.

Vaz, G. J. (2011) “E-Science na Embrapa”. Campinas: Embrapa Informática Agropecuária (Documentos, 117).