Curadoria digital e dados de pesquisa

Digital curation and research data

Luís Fernando Sayão[a]

Transcrição da apresentação

Quero dizer que é um grande prazer estar aqui na UFPB, é uma reunião muito importante porque ela analisa essa questão dos dados, ela enfoca os dados, e isso torna o evento muito importante.

A minha palestra ela tem um objetivo oculto, vou começar pela conclusão. Eu acho que a gente está tendo um grande problema no país hoje com a gestão dos dados. A gente percebe que as instituições de pesquisa e universidades começam a criar repositório de dados e esses repositórios são muito genéricos, sem especificidades, eles surgem diante de vontades políticas ou desejos de criar novos centros de informação e perdem um pouco o contato com a comunidade de pesquisadores.

Então a gente percebe em nossos estudos que é necessário que as plataformas de dados estejam muito próximas das comunidades de pesquisadores, senão a gente vai começar criar cemitério de dados, e podemos cometer o erro que cometemos com os repositórios institucionais, onde a gente percebe que não tem contato próximo com os usuários e acabam perdendo as funções. Então a gente tem a chance de não cometer esse erro novamente, de corrigir a trajetória, nós percebemos que a tendência mundial e que os repositórios nascem nas comunidades e sejam apoiados pelas bibliotecas de pesquisa e você ter um contato orgânico com os pesquisadores.

É um enfoque multidisciplinar e como colocou pretexto sobre isso, gostaria de falar um pouco sobre uma questão muito importante, é que uma parte da composição acadêmica desaparece, as nossas infraestruturas de dados não são capazes de tornar visível a produção de dados dos nossos pesquisadores.

Outra questão complicada é que nós temos que atender e tratar a curadoria dos dados uniforme, sem distinguir essas culturas, então essa palestra é tentar esclarecer um pouco esse ponto de vista de expor a necessidade de criar plataformas ligadas organicamente com as unidades de pesquisadores e que a gente enfoque o foco de gestão de dados como plataforma como múltiplos serviços e que tenha visibilidade.

Queria mostrar para vocês como é necessário a gente criar padrões específicos para isso, então esse aqui é o título da palestra e com o subtítulo aqui: a gestão do que a gente chama de cauda longa da pesquisa. Essa questão aqui é nosso alvo de estudo, e o que a gente percebe é que a gente tem uma riqueza fantástica que conta a história da geração de conhecimento, do fluxo de produção de sites, do que acontecem nesses laboratórios, novo experimento, processa os dados, planeja, faz e refaz, e toda essa trajetória de erros e acertos ela é perdida, porque só uma parte

a Centro de Informações Nucleares (CIN). E-mail: lsayao@cnen.gov.br. Currículo: http://lattes.cnpq.br/3422623122948389

pequena se torna publicados em livros, revistas, artigos científicos, e a gente chama de dados de autenticidade.

Você tem toda uma história aqui de descoberta, de coleta, de geração de dados que está oculta por trás do gráfico, e esse é um ponto importante. Você tem uma parte da ciência visível que é o que aparece nos periódicos, e uma parte invisível que a gente desconhece. Essa é a questão geral que temos aqui.

Esses dados, se eles aparecerem em plataformas de dados, se fossem tratados, poderiam gerar novos conhecimentos, novos horizontes, que é uma coisa que a gente já sabe. Eles servem para reúso de dados científicos, validar as pesquisas, ajudar com os outros pares, porque a ciência não é perfeita, ela tem um monte de deslizes, de imperfeições, fraudes, como é feita por humano ela também tem falhas. Então os dados servem para a autoafirmação da ciência, novas interpretações e basicamente nas pesquisas. Essa parte que ficou invisível se for tratada, se passar por processo de gestão, ela pode servir para essas coisas aqui.

Mas para que isso possa acontecer, para que esses processos de gestação, geração, coleta e processamento possam se tornar insumos nas nossas pesquisas é necessário que a gente crie infraestruturas tecnológicas, políticas, institucionais, para dar visibilidade a essas atividades. É um trabalho importante das bibliotecas científicas é apoiar essa visibilidade ao processo de geração de dados.

Uma coisa importante para gente é que nós temos essas infraestruturas, não só tecnológica mas variam, são políticas gerenciais e informacionais. E a gente vê que a sociedade demanda por isso, as agências de fomento querem que você faça planejamento de seus dados, faça gestão desses dados, faça planos de compartilhamento, ciência aberta, e-Science, e assim por diante.

Para o gestor científico, professor, os dados são fontes importantíssimas para a prática da ciência, na Medicina por exemplo as imagens são super importantes. Então eu preciso de infraestruturas que começam lá nas políticas internacionais e nacionais até chegar nas infraestruturas tecnológicas.

Na gestão dos dados existe isso, é uma infraestrutura que começa lá no político, quando se estabelece formas de recompensa, de fomento, de distribuição desses dados, até chegar às especificidades do documento, ferramenta, preservação, etc..

A gente tem essa ideia de que os repositórios vão ser a base de tudo, mas a gente precisa de política, financiamento, capacitação, pesquisa, legislação, parcerias, tecnologias e padrões. Uma coisa fundamental para essa gestão são os metadados, que são da área de informação. Nós temos uma parcela da parte da computação que em parte está resolvida, e uma enorme parcela de recurso de informação que devem ser colocados na gestão de dados de pesquisa.

Essa infraestrutura não está só no repositório, o repositório é um subconjunto de toda essa tecnologia que está em volta. Você precisa de recursos computacionais para modelagem, simulação, armazenamento, curadoria, colaboração, uma série de coisas que criam esse conceito de plataforma. Parece uma coisa banal mas é muito importante a gente pensar nessa situação mais ampla.

Por que gerenciar o dado? Preservar a integridade da pesquisa, permitir que os dados estejam disponíveis, e aí por diante, tem uma série de motivos para gerenciar os dados.

Aí surge uma questão fundamental, que é que nós temos a tendência de tratar os dados da mesma forma, parece que os dados são todos iguais, mas eles são heterogêneos, essa é a grande dificuldade dessa gestão. Aqui existe uma questão muito importante, a gente chama de Big Science, que são os grandes projetos com grandes instrumentos, altos custos, muitos colaboradores, por exemplo, a área de genoma, astronomia, física de partículas, o grande colisor lá na Europa que é o maior artefato humano de pesquisa, que gera 15 milhões de terabytes por ano, na parte da tarde vou falar mais sobre essa questão do Big Science.

E tem também a Small Science, que é a ciência feita no laboratório das universidades por exemplo, com pequenos instrumentos, baixo custo, quem financia isso é bolsa, pequena duração no máximo 3 anos, equipe pequena às vezes o pesquisador somente, e pesquisa local com abrangência ampla. Essas pesquisas a gente chama de posse, e trabalha para comprovar situação de posse. Na Big Science você tem algo relacionado com dados, a gente vai ver isso a tarde com mais discrição, na mesa.

Essa grande ciência não é uma coisa nova, esse cara aqui era um físico, historiador, da ciência da informação, que em 1976 já falava sobre o aspecto da grande escala da ciência moderna, nova, brilhante e curiosa, é tão evidente que foi criada para cunhar o termo grande ciência. A grande ciência é tão recente que muitos de nós não abordamos a suas origens, então a grande ciência é tão vasta que começamos a nos preocupar com o tamanho do monstro que criamos, que talvez nos gere nostalgia da pequena ciência do passado.

Mas não foi ele que cunhou esse termo grande ciência, foi esse cara aqui [leitura da citação presente no slide].

Vejam só, nessa época já se tinha essa visão de que o fomento, os recursos, eram emblemáticos para a ciência, os dois meses que eu estive em São Paulo, fui lá na FAPESP fazer uma palestra, e eu percebi que os caras já falam sobre uma Big Science, eles já se preocupam com os dados que estão nos laboratórios, e esses dados são os mais importantes que a ciência tem.

As características dessas duas áreas são super importantes para a gente pensar nos processos de gestão, por exemplo, na Big Science a gente tem uma uniformidade dos dados, são muito iguais, são gerados pelos mesmos instrumentos, então lá no CERN na Europa os dados têm o mesmo formato, são iguaizinhos, então é mais fácil gerenciar esses dados porque são parecidos. Os procedimentos de geração de dados são padronizados, mas na cauda longa desses laboratórios não, você tem gente coletando dados manualmente e são específicos, cada laboratório, cada centro de pesquisa tem uma especificidade. A gestão e a curadoria estão presentes, assim como no projeto genoma gera dados eles são curados.

No mundo da Small Science é curadoria manual, os repositórios institucionais são disciplinares, voltados para uma formação orgânica da comunidade, e na cauda longa desses laboratórios a gente tem os laboratórios multidisciplinares, porque os repositórios institucionais são muito problemáticos para fazer a gestão desses dados.

A preservação na grande ciência tem sistemas de storage, aqui a gente tem pendrive na gaveta. A grande ciência não usa esse conceito de repositório, mas são grandes bancos de dados específicos modelados especificamente para trabalhar esses dados.

A gente tem na pequena ciência basicamente planilhas, os administradores usam planilhas, etc. Vejam a grande diferença fantástica, na grande ciência eles já tem tudo, é o paraíso dos dados, o acesso é aberto na Big Science, então o cara da grande produção gera 600 milhões de fenômenos por segundo, e esses dados não podem ser analisados em um só sistema, eles são distribuídos para o mundo todo, não é uma dádiva, não é humanista, mas é uma necessidade, um imperativo desse tipo de ciência.

O reúso é imediato, o cara que tá lá em Singapura ou São Paulo recebe os dados imediatamente. Na Small o acesso é escuro, protegido, o cara bota o pendrive no bolso ou no seu computador e some, ele se torna invisível. O reúso é imediato para analisar na Big Science e na Small Science não.

E o financiamento na Big Science é fluxo contínuo, apoio internacional, etc. Na pequena ciência é por projeto. Há o reconhecimento da recompensa, o pesquisador dos grandes projetos são recompensados por gerir os dados. Aqui não, o CNPq não liga se você não fazer a gestão de seus dados.

Esses caras da Big Science tem expertise, pesquisadores, cientistas de dados, bibliotecas de dados, arquivistas, instrumentos fantásticos, satélites, colisores, sensores, recursos computacionais, software, rede, é o céu dos dados. Os meus laboratórios são o inferno, ou o purgatório dos dados.

Será que essa ciência produzida nesses laboratórios é uma ciência de segunda linha por ser menor? Não, pelo contrário. Aqui você chama de cauda longo porque a maioria dos dados é produzida pelos pequenos laboratórios, o Big Data científico não está na cabeça e sim na cauda longa.

Dados da grande ciência são fácil de vincular e arquivar, a pequena ciência é mais heterogênea e vasta, e gera duas ou três vezes mais dados. Então o Big Data científico são os dados que estão no nosso laboratório, olhem aqui, nos datasets, pequenos laboratórios, e aqui grandes projetos. Aqui são poucos projetos que geram muitos dados e ali são múltiplos projetos que geram poucos dados, são coisas pequenas, planilhas, interessante isso daqui.

Esse é o nosso problema, porque uma parte considerável disso daqui não está publicada, desaparece, some, não serve para nada.

Os dados da cauda longa são super heterogêneos, eles variam em parâmetros, tamanho, tem coleções imensas, são variados, então cada laboratório tem um formato, as vezes proprietário, mesmo os laboratórios que fazem pesquisa e geram em formatos criam uma diversidade muito grande, as estruturas são diferentes, as vezes você tem dados primários, algumas em planilhas outras em banco de dados. O nível de complexidade de dados, alguns dados simples que são imagens outros são compostos por vários arquivos e com especificidades como banco de dados por exemplo, e variam em termos disciplinares.

Se a gente olhar aqui, a cauda longa se estende por todos ramos disciplinares, domínios tecnológicos e as comunidades também, ciências humanas, ciências sociais, então a gente tem uma vasta heterogeneidade nesses termos aqui, tamanho, formato, estrutura, complexidade, tecnologia, etc.

Vocês estão vendo que a gente tem uma complexidade muito grande para fazer sistemas que possam lidar e gerenciar dados na cauda longa, que são dados produzidos pelos laboratórios no cotidiano das universidades. Esse é um problema terrível mas não é local, é do mundo todo, se procurar na Internet vai perceber que tem muita gente trabalhando com a complexidade dos dados gerados na cauda longa das pesquisas.

A gente tem que grande partes destes dados aqui da cauda longa não são publicados, grande parte não vão nem para os periódicos. Então aqui tem a revisão da literatura, isso aqui é o que é publicado e isso aqui o que não é, que é um pequeno número de projetos científicos que geram muitos dados.

Ela é pequena mas essa não é uma ciência menos importante, uma coisa interessante é que não é aqui que surge as grandes descobertas, é na cauda da ciência, aqui na grande ciência os dados já são previsíveis, não tem surpresa, não tem inovação.

Então essa área aqui se torna super importante na ciência, é aqui que surgem os modelos, as patentes, os artigos científicos, a inovação, o conhecimento multidisciplinar, e a autonomia. Tem uma frase que diz: parece que é provável que a ciência inovadora venha mais da cauda longa que da cabeça, por isso a gente tem que gerenciar esses dados.

Tem autor que diz o seguinte, a Big Science legal é a que gera muitos projetos na pequena ciência, e é verdade. Se você tem duas áreas super importantes, essas áreas são silos de dados, você precisa de dados gerados na cauda longa e dados gerados no Big Science para criar a ecologia de dados.

A nova ciência ela precisa da Big Science mas as grandes ciências são da pequena ciência. Na astronomia é bem interessante porque você tem grandes projetos e observatórios espaciais e você tem os laboratórios terrestres que fazem parte da pequena ciência, e a relação entre os dois e homogênea, produtiva, eles conseguem gerar conhecimento entrelaçando os dados da cauda longa com a cabeça, então essa ecologia dos dados é produtiva.

Esse texto aqui, a perspectiva sistema espaço – dado torna-se chave para as respostas na nova ciência, isso acontece ao vincular a estabilidade da grande ciência ao território de autonomia da cauda longa, cuja conduta desses criadores leva a inovação e geração de conhecimento multidisciplinar.

A gente precisa das duas ciências, essa é a questão. E que as coisas importantes inclusive a geração de empregos necessitam que a ciência seja feita nas nossas universidades, e a gente esquece que a riqueza desses projetos são eles que geram os artigos em periódicos, não é a grande ciência que gera artigos em periódicos.

Então a gente tem essa pirâmide aqui que eu que fiz uma adaptação, então aqui você tem os recursos relevantes, banco de dados maiores, grande colisor, definição de formatos, e assim por diante, esses caras aqui são super organizados. Depois a gente tem os repositórios disciplinares, eles são super entrosados, tem princípios da comunidade, da disciplina, só de modelos biológicos, então eles nascem da comunidade e crescem em sistemas específicos. Então na análise de genoma você tem vários bancos de dados que se juntam.

Depois é como se estivesse caindo no inferno aqui, o inferno são os repositórios multidisciplinares e institucionais, que não é capaz de oferecer serviços de curadoria, e o acesso é

muito restrito, e tem uma coisa muito importante, a gente fala muito da indexação mas pouca gente fala sobre a manipulação dos dados, alguém já ouviu falar sobre recuperação de dados. Essa questão nos repositórios multidisciplinares você não consegue recuperar os dados que você quer, você recuperar grandes coleções ou segmento específico, então não tem metadados para poder lidar com isso tudo, não tem serviço aos pesquisadores por isso que eu acho que os repositórios de dados multidisciplinares e incluem os institucionais podem virar cemitério de dados.

Aqui por último estão as coleções individuais, aqui é o purgatório e aqui é o inferno, e aqui é o céu, a gente precisa criar o céu para os dados da cauda longa. O sucesso na prestação de serviços para gestão de dados de pesquisa está relacionado a sua capacidade de dar apoio as classes e culturas da comunidade científica da instituição, por isso não adianta criar grandes repositórios de dados porque senão eles podem virar cemitério de dados.

Será que a gente está construindo cemitério de dados em repositórios? Que você vai guardar bases de dados mal apresentados, não vão ser recuperados, não vão ser preservados, não vai ter curadoria? Essa é a questão, então a solução pode ser que a gente tenha que construir repositórios talhados a cada área, disciplina, comunidades, e depois crie a interoperabilidade. Mas também a gente não pode criar silos, cofres, a gente tem que fazer com que esses repositórios se falam via Web Semântica, Ontologias, e a gente quer também o reúso.

Outra coisa importante nesse mundo que a gente está trabalhando é saber por que a gente está em um novo partido. A maioria dos pesquisadores concordam com os principais meios de compartilhamento e reúso, mas relutam a participação dos próprios dados.

Por que isso? Em termos gerais o cara tem culturais, tem disciplina que compartilhar faz parte, como exemplo a Astronomia, Física, mas tem outras que o dado é um segredo, Química por exemplo é super fechado, a área nuclear em que eu trabalho. Então isso também faz parte da cultura disciplinar, para a Astronomia é fundamental.

Interesses econômicos, comerciais, por exemplo.

Resultados negativos o cara acha que se não for confirmado ou se der errado ele não vai publicar os dados mas ele tem que publicar. Hoje a gente tem periódicos de dados negativos, é tão importante quando você ter um novo tipo de periódico que se dedica isso.

Custos, é muito custoso em termos de tempo e dinheiro você tratar os dados, é muita coisa, criar e limpar os dados, processar, publicar, é muito complicado, então ele prefere publicar em periódico.

Vantagem produtiva, o cara tem que publicar o máximo que ele pode, publicar em periódico o máximo que ele puder, depois abona os dados, coloca no pendrive e não compartilha. Medo muito grande dos dados serem erroneamente interpretados, e a dificuldade na área de Medicina por exemplo, o cara tem muito medo disso.

E a gente foi mais adiante nisso e trabalhou numa pesquisa de motivos mais específica, e 50% não são publicados, e a gente viu motivos individuais como atitude dos caras, disciplinares, organizacionais e políticas.

Uma coisa importante é poder ver as agências de fomento e as revistas que exige para que você publique seus dados.

E a gente essas políticas que começam lá em cima, requisitos internacionais, vai descendo até o nível disciplinar, essas coisas vão se ajustando às comunidades específicas.

Aqui eu tenho os motivos pessoais que vão de oportunismo a medo de lançar os dados que revela o que foi feito, e assim por diante. Tem uma série de motivos que o pesquisador da cauda longa não publique seus dados, por isso nós temos que criar infraestrutura para coordenar isso.

A gente não tem que focar em repositório, repositório é um subsistema de um sistema mais amplo que faz serviços mas que precisa também se preocupar com a visibilidade do pesquisador, a citação, o conhecimento, criar a memória dinâmica da instituição, criar serviços inovadores, a gente percebe que os repositórios têm serviços muito talhados nos dados, serviços bacanas, então grande problema que a gente tem hoje é que não tem serviços, tem uma submissão e um acesso que é muito precário.

Tanto é que agora o esforço para se criar maior infraestrutura de repositório que tenha mais serviços, mas isso só é possível se estiverem próximos das comunidades e disciplinas.

O pesquisador precisa de capacitação, então a ideia da plataforma é mais ampla que a ideia do repositório, simplesmente um mecanismo incluído aqui.

Uma coisa importante nessa complexidade da gestão dos dados é que os dados não é útil quando coletados, não é como uma tese ou um livro, os dados vão se transformando, dados científicos, dados brutos, processamento, análise, e esse fluxo tem duas coisas, ele é complexo e precisa ser registrado para saber a proveniência, o histórico dos dados com metadados específicos.

E esse fluxo não é um fluxo padronizado, cada disciplina, cada laboratório, tem um tipo de fluxo, você tem que trabalhar o fluxo de cada área no laboratório, de cada linha de pesquisa, a gente tem essa complicação.

É necessário compartilhar muito mais que os dados finais, eu preciso compartilhar trajetórias de erros e acertos. A gente tentou botar aqui um fluxo padrão incluindo questões de publicação em data papers, análise, mas é uma fixação nossa, para entender como funciona, e aqui uma intervenção, é legal esse exercício para fazer o embaralhamento de um fluxo genérico.

Dados são objetos complexos que precisam ser preservados, e essa complexidade é crescente, tem várias camadas que você tem que preservar. E tem os dados confiáveis, que está se trabalhando muito agora. E uma coisa fundamental aqui é que a gente precisa dos metadados, os dados não falam por si próprio, se você tem um livro, uma tese, ou um artigo está escrito o que ele é, mas os dados às vezes é só um monte de número, monte de letras, precisam ser documentados, informação de representações, significados, semântica, estrutura.

E esse aqui é um cara aqui é Grey, cara da Informática que criou os esforços da e-Science, e era da Computação mas achava que a Biblioteconomia iria resolver essas questões todas, é muito interessante. Ele fala o seguinte, dados de pesquisa são incompreensíveis e portanto inúteis a menos que haja uma descrição detalhada de como e quando eles foram obtidos. Então eu tenho que detalhar com metadados todas essas coisas aqui.

A gente tem um negócio que chama metadados similares, porque cada área é específica, e esses metadados documentam o processo todo, resolução, limpeza, tudo mais.

Os dados são versionados, e tudo isso para o reúso, mas o reúso também é uma coisa complexa. As atividades que são locais tem que se tornar globais e explícitas, então você pegar o conhecimento do seu laboratório e tomar como base, é a complexidade.

Tem um monte de coisa que você tem que fazer além da propriedade tecnológica, tem um grau muito grande a ser levado em conta. A gente tem que ter os dados FAIR mas a gente acha melhor os dados "manerios", em São Paulo é "daora", existem outros conceitos de FAIR.

A gente tem que contornar isso incentivando capacitação e financiamento, hoje a gente tem um monte de outras formas de publicação, dados negativos publicados em repositórios específico, publicação ampliada, e nos laboratórios que colocam da Internet.

A gente tem que sair daquela plataforma genérica que é stop down, que é da reitoria, e sem preservação, sem serviço, ir para plataforma disciplinar, serviços disciplinares, cursos, a gente tem que tentar, obrigado.

Vídeo da apresentação

Título: Curadoria digital e dados de pesquisa.

Disponível em: http://dadosabertos.info/enhanced_publications/idt/video.php?id=30

Slides da apresentação

Título: Curadoria digital e dados de pesquisa.

Disponível em: http://dadosabertos.info/enhanced_publications/idt/presentation.php?id=30