2014-01-16

Digitalização e preservação digital: a experiência do Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP)

DIGITALIZAÇÃO E PRESERVAÇÃO DIGITAL: a experiência do Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP)

Sueli Mara Soares Pinto Ferreira[1]
Zacharias Gadelha[2]
Camila Gamba[3]

RESUMO
As atividades de digitalização e de preservação digital de conteúdos permeiam as ações das bibliotecas universitárias, quer seja para a preservação de acervos raros e especiais como para garantir o acesso amplo à produção intelectual da Universidade. São descritas as experiências desenvolvidas pelo Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP) referente as ações implementadas em âmbito sistêmico, abordando a construção da infraestrutura tecnológica (hardware e software) e as políticas e normas institucionais de digitalização e de preservação digital, apresentando ao final os desenvolvimentos já obtidos com os conteúdos produzidos até o momento.
Palavras chave: Digitalização, Preservação Digital, Biblioteca acadêmica, Biblioteca Digital

ABSTRACT

The digitization and digital preservation activities are presented in all the academic libraries actions today. This is done for the preservation of rare collections and special collections and also to ensure wide access to the intellectual production of the University. Here we have described the experiences developed by the University of Sao Paulo Integrated Library System (SIBiUSP) in reference to the systemic actions implemented, describing the technological infrastructure construction (hardware and software) and the institutional policies and standards for digitalization and digital preservation, presenting at the end the accomplishments with the content up to this date.

Keyword: Digitization, Digital preservation, Academic Library, Digital Library

1 INTRODUÇÃO

As atividades de digitalização de conteúdos em bibliotecas universitárias se tornaram contínuas e sistemáticas há algum tempo, quer seja para a preservação de acervos raros e especiais, ou para garantir o acesso amplo à produção intelectual da Universidade aumentando, assim, sua visibilidade e acessibilidade. Esse último objetivo foi ainda mais valorizado com o crescimento do movimento de acesso aberto e surgimento dos repositórios institucionais.

[1] Professora titular da FFCLRP e Coordenadora do SIBiUSP, período 2010 a 2014. Email: smferrei@usp.br
[2] Analista de Sistema e Chefe da Divisão de Tecnologia de Informação do SIBiUSP. Email: zacha@usp.br
[3] Bibliotecária e Chefe da Divisão de Gestão de Projetos do SIBiUSP. Email: camilamg@usp.br

No entanto, à medida que aumenta a produção de coleções digitais, maior é a necessidade de se estabelecer políticas específicas, tanto no que concerne à digitalização, preparação e indexação desses conteúdos como, e principalmente, quanto às questões de segurança, backup e preservação digital.

É dentro de tal direcionamento que o Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP) tem planejado suas atividades, laboratório de digitalização, políticas institucionais e infraestrutura de hardware e software. Relatam-se aqui os resultados obtidos até o momento, projetos e planos em curso.

2 INFRAESTRUTURA TECNOLÓGICA

No que se refere à digitalização, foi inaugurada em 2008, com apoio Agência Brasileira da Inovação (FINEP)[4], a Oficina de Digitalização do SIBiUSP que, naquele momento contava com apenas uma Câmera Digital Nikon D-2XS dotada de uma objetiva Nikon Micro AF 60MM e outra AF Nikkor 24-85MM F2.8-4D. De maneira ainda incipiente, iniciou-se a discussão sobre tal processo no âmbito do SIBiUSP e da própria Universidade, bem como os experimentos com a digitalização de alguns livros raros.

Em 2011, com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)[5], inicia-se a migração da antiga oficina para o atual Laboratório em Digitalização (mais focado em pesquisas e experimentações sobre o tema) que então recebe uma ampliação considerável com a aquisição de equipamentos de como o robô Kabis III que opera em alta velocidade, e o Skyview, sistema de digitalização para grandes formatos, ambos produzidos pela Kirtas Technologies, USA. O primeiro modelo veio equipado com duas câmeras Canon de 21 megapixels, dispostas em x (cada uma virada para uma página do livro), portanto com potencial para digitalização de livros encadernados. Digitaliza até 2900 páginas por hora, como utiliza duas câmeras, captura duas páginas simultaneamente. Possui viragem das páginas dos livros por braço robótico de vácuo e impondo um ritmo de trabalho constante e muito superior a qualquer tipo de digitalização manual.

[4] Projeto FINEP PROINFRA 01/2005 “Oficina de Digitalização de documentos: preservação e difusão dos acervos raros e/ou especiais da USP”.

[5] Projeto FAPESP 2009/54784-7, “Infraestrutura para a pesquisa de coleções raras e especiais da USP/UNESP/UNICAMP: recolhimento, preservação, organização e disponibilização para acesso à comunidade científica nacional e internacional”. Projeto aprovado em 2010 e iniciado em 2011, ainda em andamento.

Já o modelo Skyview (planetária) é voltado para a digitalização de grandes formatos como mapas, cartazes e jornais, mantendo boa resolução e qualidade. Possui apenas uma câmera que se desloca em dois eixos para varrer toda a extensão do material. Cada câmera é ligada num computador que, por sua vez, é ligado a um servidor. As imagens aparecem em tempo real no monitor do scanner. O disparo da câmera pode ser automatizado, impondo-se assim um ritmo constante ao trabalho, ou pode ser por pedal, para uso com obras mais delicadas e conforto ao operador. Possui uma mesa de vácuo para garantir que a página estará sempre sem ondulações e possa gerar uma imagem fiel, contando com ajuste de sucção. Um berço com compensação de altura e lombada, o que permite o uso de um vidro planificador com o mínimo de estresse sobre a costura ou as páginas das obras.

Ainda em 2012, novamente com apoio da FINEP, foi adquirido o Scanback modelo Archive, com sensor de 312 MPix, 48bits trilinear, da empresa alemã Rencay, de altíssima resolução e fidelidade para objetos em diversos tamanhos e formatos. Considerado Scanback para área artística, sendo cada cor capturada de forma independente e sem interpolação, e cada ponto capturado "três vezes", produzindo um resultado final extremamente fidedigno à realidade do objeto capturado. Utiliza sensores com o dobro de bits que os profissionais, além de possuir um sensor bem maior do que o das câmeras convencionais com resolução em média 10x maior que o comum das câmeras profissionais.

Em síntese, o Laboratório de Digitalização do SIBiUSP, hoje, possui condições de digitalizar tanto formatos padrão quanto grandes; realiza digitalizações de alta qualidade com alta produtividade, possibilita capturas de real preservação e com qualidade artística considerável, tem condições de operar documentos de dimensões de até três metros de largura mantendo-se ainda a alta resolução e fidelidade.

Para sustentar o desenvolvimento de todos esses projetos anteriormente descritos, ao mesmo tempo em que se observava o foco principal da atual gestão do SIBiUSP centrado na digitalização, indexação e preservação da produção USP, foi feito em final de 2010, um criterioso mapeamento da quantidade de documentos publicados pela comunidade USP (especialmente teses, dissertações e revistas científicas), em formato impresso armazenados nas setenta e três bibliotecas das USP, alocadas em unidades dispersas em sete cidades paulistas

Somando-se o acervo de revistas ao de teses, totalizaram 169 mil documentos em formato impresso, correspondendo a aproximadamente 39 milhões de páginas de conteúdo a espera de serem digitalizadas e disponibilizadas online. Segundo nossas análises, esse total de páginas, quando digitalizadas, vai gerar 706TB de arquivos TIFF/RAW para armazenamento e 265TB de arquivos em PDF para consulta pelos usuários finais.

Com base nesse contexto, a equipe do SIBiUSP identificou três novas situações problemas: (1) como evitar imensos custos de transporte do material a ser digitalizado, minimizar o risco de extravio, de danos no material e diminuir ainda o tempo de interrupção de acesso aos exemplares físicos nas bibliotecas de origem? (2) como e onde armazenar todo esse material digitalizado ou nascido digital? (3) como garantir a preservação e longo prazo aos documentos USP digitalizados?

A resposta à primeira questão resultou na proposição de novo projeto de infraestrutura na FINEP[6] aprovado em 2012 e iniciado esse ano de 2013, para a construção de oficinas móveis de digitalização, ou seja, oficinas em containers que serão deslocados diretamente para o local onde está armazenado o conteúdo físico a ser digitalizado. Portanto, se encontra agora em fase de implementação de quatro oficinas (ramais do Laboratório de Digitalização do SIBiUSP) que possam atender à Universidade, visto que poderão circular rotativamente entre suas unidades de ensino e pesquisa para permitir o processo de migração digital diretamente no local físico onde se encontra a documentação científica e acadêmica. Tais oficinas irão permitir a digitalização, tratamento e disponibilização na web do conteúdo completo das revistas cientificas publicadas por essa Universidade, desde a primeira metade do século XX, e das teses e dissertações defendidas, no âmbito da USP, desde os primórdios de seus diversos programas de pós-graduação.

As questões dois e três levaram ao planejamento da infraestrutura de hardware e software necessária para o desenvolvimento de uma política de preservação digital completa, iniciando-se pelas necessidades de armazenagem imediata. Especificamente para esse projeto das oficinas móveis, foi concebida uma infraestrutura com capacidade de armazenamento estimada inicialmente em 1,6 Petabytes de dados. Com esta capacidade, estimou-se ser possível armazenar, nos diversos campi da Universidade, versões formatadas em padrão PDF/A dos arquivos digitais produzidos (para acesso via web) e manter armazenados maciçamente no campus central em São Paulo, as versões em RAW ou TIFF dos mesmos conteúdos digitalizados (para preservação e respectivos backups). Desta forma, já foram adquiridos os seguintes equipamentos NETAPP:

(a) um *storage* para o armazenamento central com capacidade bruta de 1,72 PB que deverá estar fisicamente alocado no campus principal da USP na cidade de São Paulo;

(b) quatro *storages* com 48 TB para montar postos de armazenagem distribuídos nos campi do interior, com o propósito de armazenar o material em processamento nas oficinas móveis, bem como os PDF finais que ficarão disponíveis para consulta dos usuários. Também se responsabiliza pela redundância do material disponível. A logística de distribuição e implementação desses centros nevrálgicos se encontra em fase de

[6] Projeto MCT/FINEP/CT-INFRA – PROINFRA – 01/2011 – Plano Plurianual de Infraestrutura em Pesquisa da USP: Tecnologia de Informação e Core Facilities (Fase II)

conclusão, mas certamente o arranjo será determinado de acordo com a qualidade e largura de banda de rede local.

(c) um *storage* intermediário, de 96 TB, para integração local, backup e DR (disaster recover) da produção. Sediado em São Paulo, deverá ser o gerenciador de documentos em todos os outros, verifica distribuição dos lotes entre os servidores dos campi.

(d) uma máquina de "ingestão", como está sendo chamada, preparada para receber a produção digitalizada (antes do tratamento), cuidar dos backups, recuperação. Com esse equipamento, é possível efetuar, automaticamente, a captura de conteúdos já deduplicados e comprimidos (compressão) previamente nas próprias máquinas remotas.

(e) cinco servidores físicos Itautec, um para cada uma das localidades (São Paulo e interior), que atuarão como estação de controle e servidor de acesso ao conteúdo armazenado nos storages anteriormente mencionados.

Todos estes equipamentos já contêm os softwares e licenças necessários para integração, backup e replicação para outras estruturas USP e incluem, também, softwares para *compliance* (controle do cumprimento de regulamento de retenção e proteção de registros e documentos).

Uma das premissas, desde o início do Programa SIBiUSP de digitalização e preservação digital, foi sua total aderência ao projeto institucional "CloudUSP" (computação em nuvem) vinculado a atual Vice-Reitoria de Administração. Esse projeto, iniciado em 2010, é viabilizado por meio de estruturas de computadores, contendo milhares de servidores de aplicações e grandes *storages* para armazenamento de conteúdo, agrupados e organizados em *Internet Data Centers* (IDCs) na capital e nos diversos campi, interligados à Internet de alta velocidade. Em 2013, passa a disponibilizar, servidores virtuais para as suas Unidades de informática (FONTE revista Espaço Aberto), visando a racionalização dos gastos com aquisições para tal infraestrutura, consequentemente, reduzindo custos diretos e indiretos de manutenção (licenças, energia, segurança, recursos humanos especializados etc).

Finalmente, dando continuidade às definições de infraestrutura para o SIBiUSP e preocupados não somente com a preservação de todos esse conteúdo que estará sendo digitalizado, mas também com aqueles que já estão nascendo em formato digital e ocupando espaço nas bibliotecas digitais da USP, uma próxima rodada de ações envolve a aquisição de software para garantir a operacionalização da política de preservação digital e o workflow do processo de criação e gestão de conteúdo, desde autenticidade, integridade dos arquivos, entrega para sistemas de busca, visões personalizadas de acordo com filtros específicos internos e externos. Se aprovado esse projeto, também submetido à FINEP, para 2013 será iniciada sua primeira com à análise dos softwares disponíveis no mercado, a existência de padrões e normas internacionais bem como sua adequação à realidade USP.

3 POLÍTICAS INSTITUCIONAIS

Desde 2008, por meio da Portaria GR no. 4035, foi criada a Comissão de Digitalização das Obras Raras e Especiais das Bibliotecas do Sistema Integrado de Bibliotecas da USP com o objetivo de estudar, planejar e estabelecer diretrizes para a digitalização desse tipo de documento.

Atualmente, tal portaria se encontra em fase de atualização, visando incluir não apenas obras raras, mas, especialmente, a produção científica da USP. Do mesmo modo, estão finalizadas as primeiras diretrizes para ações dessa natureza, de modo a embasar a Universidade e sua comunidade, de orientações mínimas e essenciais a serem seguidas no processo de digitalização, no tratamento das imagens, na utilização de OCR, na produção de arquivos para usuários (seja em PDF, PDFA, dispositivos móveis etc), na indexação em bibliotecas digitais com base em metadados normalizados , projetos e normas de preservação nacionais e internacionais (tais como CONARQ, DFG, OAIS dentre outros).

Outro estudo em finalização é a política de preservação digital do SIBiUSP que deverá determinar claramente o que deve ser armazenado, como priorizar e organizar conteúdos e, finalmente, gerar o fluxo de armazenamento em diferentes mídias e locais segundo a infraestrutura descrita anteriormente.

No que se refere a backup e redundância de dados, o SIBiUSP é um nó na Rede Cariniana mantida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Iniciando com um piloto no Portal de Revistas USP, já foram implementados a instalação do servidor, teste de comunicação entre as unidades participantes, teste de depósito dentre outras atividades. O objetivo do SIBiUSP em participar dessa rede é atuar em âmbito nacional, disponibilizando e divulgando o conteúdo USP, compartilhando experiências com outras universidades e apoiando o desenvolvimento do tema no país.

Do mesmo modo, está em estudo a implantação de um sistema distribuído de preservação digital próprio para o SIBiUSP, utilizando-se uma rede LOCKSS privada (PNL) com apoio direto da Universidade de Stanford e do próprio IBICT, visando proceder à redundância dos próprios dados USP, melhorando a performance do sistema e a recuperação de informações pelo usuário.

4 CONTEÚDOS DIGITALIZADOS e BIBLIOTECAS DIGITAIS

O SIBiUSP lançou em 2003 a sua primeira Biblioteca Digital de Obras Raras e Especiais, com o objetivo de garantir a salvaguarda, a disseminação e o acesso aos acervos bibliográficos da Universidade de São Paulo, e com o apoio do CNPQ digitalizou e disponibilizou 38 obras raras em alta resolução.

Em 2011, ainda com o projeto FAPESP 2009/54784-7, além da atualização dos equipamentos do Laboratório, ocorreu a necessária revisão dos critérios de digitalização de conteúdos USP, da normalização internacional dos padrões de metadados a serem seguidos e o replanejamento da própria biblioteca digital que abrigaria tais conteúdos.

Como parte do processo de digitalização definido, as matrizes das imagens são geradas em alta resolução, gerando arquivos de alta definição (armazenado para preservação). Já os arquivos de acesso pelos usuários finais são processados por meio do software BSE (Book Scan Editor da própria Kirtas), ajustando-se parâmetros como brilho, contraste e remoção de manchas. Atualmente, se encontra em teste o software gratuito *Scan Tailor*[7].

O reconhecimento ótico de caracteres (OCR - *Optical Character Recognition*) também é feito recorrendo-se a software proprietário da Kirtas, resultando em um arquivo PDF que agrupa as imagens da digitalização e o texto final. Por fim, utiliza-se o software LuraDocument® PDF Compressor para compactação dos arquivos PDF.

Durante o procedimento de digitalização, diversos metadados administrativos e técnicos são gerados automaticamente em linguagem de marcação *xml* para cada arquivo digitalizado. Esses metadados, e outros técnicos definidos com base em diversos projetos de digitalização e preservação digital, enriquecem o padrão básico do Dublin Core.

Até o momento já foram digitalizados mais de 2000 obras (livros e revistas) raros e especiais na área de ciências biológicas. Encontra-se em processo a digitalização da coleção completa de jornais publicados na cidade paulista de Itú no século passado (1873 a 1961), são eles: “República”, "O Ytuano" e "Imprensa Ytuana". Também se encontra em processo a digitalização das mais antigas e raras obras selecionadas da Coleção Cervantina doada à USP recentemente, contemplando edições diversas do livro Dom Quixote em inúmeros idiomas.

[7] http://scantailor.sourceforge.net/

No que se refere à produção da Universidade de São Paulo, conforme já mencionado, enorme esforço tem sido feito para digitalizar a coleção completa de suas revistas. Nesse sentido, serão lançadas ao público em novembro desse ano, como parte das comemorações centenárias, as coleções integrais da Revista de Medicina e O Bisturi da Faculdade de Medicina da USP, e a coleção da Revista de Direito " da Faculdade de Direito da USP que está completando 120 anos.

No que se refere ao ambiente digital para abrigar tal conteúdo, optou-se também por migrar do ambiente *in-house* da primeira biblioteca digital para o software DSpace devido a uma série de necessidades, em especial o uso de uma tecnologia já dominada pela equipe do próprio SIBiUSP, pelas características de preservação digital que apresenta, mas principalmente pela possibilidade de se estabelecer um fluxo de trabalho que permitisse alimentação descentralizada, e ao mesmo tempo coordenada, pelas distintas equipes das bibliotecas USP.

Hoje, a Biblioteca Digital de Obras Raras e Especiais da USP[8] utiliza a versão 3.1 do software DSpace e a interface Corisco desenvolvida pela equipe do projeto Brasiliana Digital USP. Possibilita a busca por tipo de documento, autor, títulos, imagens especiais, biblioteca detentora da coleção física, além de oferecer textos de curadoria específica buscando evidenciar diferentes perspectivas das coleções digitais ali armazenadas, facilitar a busca e motivar o interesse dos usuários. Está preparada para oferecer à comunidade USP, produtora de conteúdos digitais referentes a acervos históricos, raros e especiais resultantes de seus projetos de pesquisas, bem como as equipes das bibliotecas USP o armazenamento e gerenciamento de seus arquivos digitais. Além dessa Biblioteca, conteúdos digitais referente as teses e dissertações são dirigidos à Biblioteca Digital de Teses e Dissertações da USP[9], a produção intelectual da USP (científico, acadêmica, artístisca e técnica) são direcionadas a Biblioteca Digital da Produção Intelectual da USP[10] e as revistas correntes são indexadas no Portal de Revistas USP[11].

[8] http://www.bore.usp.br
[9] http://www.teses.usp.br
[10] http://www.producao.usp.br
[11] http://www.revistas.usp.br

5 CONSIDERAÇÕES FINAIS

A criação de uma infraestrutura tecnológica descentralizada, a proposição de políticas institucionais de digitalização e de preservação digital e a oferta de um ambiente de biblioteca digital aberto para gestão descentralizada de conteúdos pela própria comunidade uspiana compõem o Programa de Digitalização e Preservação Digital em curso atualmente no SIBiUSP.

Embora ainda em fase de consolidação, o impacto de tais proposições já se faz sentir por meio da crescente procura da expertise de nossas equipes pelas demais unidades, institutos e museus USP visando estabelecer parcerias e buscar apoio no desenvolvimento de distintos projetos, envolvendo desde digitalização, até indexação de conteúdos e uso da infraestrutura da biblioteca digital de obras raras e especiais para gestão de suas coleções.

No entanto, faz-se urgente a formalização e operacionalização da política de preservação digital, de modo a garantir a necessária e criteriosa salvaguarda do acervo USP, a ampla e irrestrita acessibilidade aos conteúdos disponíveis e a recuperação ágil e consistente de conteúdos. Portanto, a consolidação do Laboratório de Digitalização do SIBiUSP como espaço de pesquisa e referência na área, a montagem das oficinas móveis ramais, a conclusão das políticas institucionais e a implantação de um sistema especifico de preservação digital no próximo ano são pontos fundamentais e decisivos.

BIBLIOGRAFIA

A USP na era da computação em nuvem. **USP Destaques**, São Paulo, n. 66, p. 01-02, 12 set. 2012. Disponível em: <http://www.usp.br/imprensa/wp-content/uploads/USP-Destaques_66.pdf>. Acesso em: 01 set. 2013.

CONSELHO DE REITORES DAS UNIVERSIDADES ESTADUAIS PAULISTAS/BIBLIOTECAS. **Infraestrutura para a pesquisa de coleções raras e especiais da USP/UNESP/UNICAMP: recolhimento, preservação, organização e disponibilização para acesso à comunidade científica nacional e internacional**. São Paulo, 2009. Projeto FAPESP, Proc. Nº 2009/54784-7.

CONSELHO NACIONAL DE ARQUIVOS – CONARQ. **Recomendações para Digitalização de Documentos Arquivísticos Permanentes**, 2010. Disponível em: <http://www.conarq.arquivonacional.gov.br/media/publicacoes/recomenda/recomendaes_para_digitalizao.pdf>. Acesso em: 01 set. 2013.

NUVEM USP inicia segunda fase com a criação de servidores virtuais **USP Destaques**, São Paulo, n. 70, p. 01-02, 29 jan. 2013. Disponível em: <http://www.usp.br/imprensa/wp-content/uploads/Destaques-70.pdf>. Acesso em: 01 set. 2013.

STANFORD UNIVERSITY. **What Is LOCKSS?** Disponível em: <http://www.lockss.org/about/what-is-lockss/>. Acesso em: 01 set. 2013.

UNIVERSIDADE DE SÃO PAULO. **Portaria GR Nº 4035, de 01 de Dezembro de 2008.** Disponível em: < http://citrus.uspnet.usp.br/sibi/Portaria-Resolucao/port_gr_4035.htm >. Acesso em: 01 set. 2013.

UNIVERSIDADE DE SÃO PAULO/SISTEMA INTEGRADO DE BIBLIOTECAS. **Oficina de Digitalização de documentos: preservação e difusão dos acervos raros e/ou especiais da USP**. São Paulo, 2005. Projeto MCTI/FINEP/CT INFRA – PROINFRA 01/2005.

UNIVERSIDADE DE SÃO PAULO/SISTEMA INTEGRADO DE BIBLIOTECAS. **Plano Plurianual de Infraestrutura em Pesquisa da USP: Tecnologia de Informação e Core Facilities (Fase II): subprojeto OFMOVEL.** São Paulo, 2011. Projeto MCTI/FINEP/CT INFRA – PROINFRA – 01/2011.

AGRADECIMENTOS

A equipe do Programa de Digitalização e Preservação Digital do SIBiUSP, merecedora de reconhecimento pelo trabalho que vem executando, é composta por: Allan da Silva, André Nito Assada, Claudio Roberto Ferreira, José de Souza Araújo, José Luiz Gomes da Costa, Laucivaldo Cardoso, além dos autores desse relato.