LEONARDO DALLA BERNARDINA SANTOS

O PROCESSO DE IMPLANTAÇÃO DE UM REPOSITÓRIO DIGITAL DE INFORMAÇÕES BASEADO EM SOFTWARE LIVRE

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras, como parte das exigências do curso de Pós-Graduação *Lato Sensu* em Administração de Redes Linux, para a obtenção do título de especialista em Redes Linux

Orientador
Prof. Joaquim Quinteiro Uchôa

LAVRAS
MINAS GERAIS – BRASIL
2006

LEONARDO DALLA BERNARDINA SANTOS

O PROCESSO DE IMPLANTAÇÃO DE UM REPOSITÓRIO DIGITAL DE INFORMAÇÕES BASEADO EM SOFTWARE LIVRE

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras, como parte das exigências do curso de Pós-Graduação *Lato Sensu* em Administração de Redes Linux, para a obtenção do título de especialista em Redes Linux.

APROVADA em 30 de Abril de 2006

Prof. Joaquim Quinteiro Uchôa (Orientador): _____

Profa. Kátia Cilene Amaral Uchôa: _____

Prof. Denilson Vedoveto Martins: _____

LAVRAS
MINAS GERAIS – BRASIL

DEDICATÓRIA

Aos meus pais,
Fontes de inspiração e força.
Se eu chegar a ser metade do que
significam para mim, me dou por satisfeito.

AGRADECIMENTOS

Ao meu orientador, Prof. Joaquim;
me impressiona como alguém tão ocupado
pode ser sempre tão disponível.

À Profª Eliethe, minha “mãe adotiva”, por
ser sempre mais do que espero dela.

À Ellen, minha namorada. Sempre presente
na medida certa. Mais que grato a você, sou
apaixonado por você.

A Deus; por tudo, simplesmente.
Por colocar na minha vida pessoas como
essas e tantas outras. Se fosse agradecer a
todas, ocuparia mais páginas que o trabalho.

RESUMO

A instituição onde este trabalho foi desenvolvido sente já há algum tempo a necessidade de disponibilizar parte do conteúdo que produz de forma *online*, visando principalmente à preservação a longo prazo, à interoperabilidade com outras instituições e a proporcionar acesso fácil a esse material. Existem vários programas que permitem que esse objetivo seja alcançado.

Este trabalho avalia algumas das ferramentas disponíveis baseadas em *software* livre, tendo em vista as necessidades da instituição onde o trabalho foi desenvolvido e verifica se o *software* livre pode adequar-se a esse nicho de mercado que são os repositórios digitais da informação e que não se restringe apenas a bibliotecas, mas a qualquer instituição que produza informação que deva ser preservada, compartilhada e facilmente acessada.

SUMÁRIO

Lista De Figuras

LISTA DE TABELAS

1. INTRODUÇÃO

A função básica de uma biblioteca é tornar o acesso à informação mais rápido, eficiente e democrático. Isso por si só já justifica a necessidade de uma biblioteca virtual ou, num sentido mais amplo e que será usado neste trabalho, um repositório digital de informações. A necessidade mais urgente da biblioteca de ensino superior onde foi realizado este trabalho é disponibilizar o texto integral dos trabalhos acadêmicos para consulta *online*, uma vez que os exemplares impressos não são emprestados por ser difícil obter uma nova cópia em caso de extravio.

O número de instituições de ensino superior cresceu significativamente nas últimas décadas. Geralmente é do interesse das bibliotecas dessas instituições compartilhar sua produção intelectual por diversos motivos, como receber em troca a produção de outras instituições, promover a instituição para seus clientes em potencial ou mesmo simplesmente para facilitar o acesso à informação à comunidade interessada.

A maioria dos usuários dessas bibliotecas já está habituada ao uso da Internet e à agilidade e possibilidades de auto-atendimento que ela oferece e esperam isso também da biblioteca. Segundo (FURLAN, 1993):

> ”*cada vez mais os clientes estarão estruturados e exigentes, esperando receber valor agregado que a informação pode fornecer aos produtos e serviços...por muito tempo corremos atrás da tecnologia – chegou a vez da tecnologia correr atrás de nós”*.

O objetivo disso é tornar o acesso do usuário aos serviços da biblioteca o mais rico e autônomo possível.

É necessário considerar também que a instituição produz material nos demais setores, não apenas naqueles diretamente envolvidos com o ensino. Esse material pode ser usado como apoio ao desenvolvimento profissional nesses setores ou ainda guardado para fins de memória histórica. Por isso, muitas vezes acaba sendo encaminhado à biblioteca ou, se não, se perde com o tempo. Muito desse conhecimento já é originalmente digital, enquanto outra parte é digitalizada posteriormente. De qualquer forma, essas obras são candidatas a ajudar a compor o acervo de um repositório digital de informações.

Considerando tudo isso é que a instituição onde este trabalho foi desenvolvido decidiu investir na implantação de um repositório digital de informações disponível via *web*. É importante observar que o termo "repositório digital de informações" também pode ser encontrado na literatura como "biblioteca digital", "biblioteca virtual" ou "base de dados de informações" e que em alguns casos eles são utilizados de forma intercambiável. Neste trabalho será utilizado preferencialmente o termo "repositório digital de informações", referindo-se a um *software* instalado, disponível e com conteúdo que possa ser acessado, ou seja, a todo o conjunto de *software,* informações armazenadas, procedimentos, equipamentos e pessoas que disponibiliza conhecimento ao usuário final. Quando tratar-se apenas do *software* que permite armazenar e recuperar informações será usado só o termo "*software*".

Neste contexto o termo "obra" refere-se a qualquer registro de informação. A informação pode estar registrada em diferentes formatos

(áudio, vídeo, texto) e suportes (fita K7 ou CD, DVD ou fita VHS, livro ou jornal, respectivamente). Uma obra, como será tratada aqui, pode assumir qualquer formato ou mesmo mais de um formato simultaneamente. Por exemplo, um livro pode conter anexo um CD-ROM; ambos serão considerados, em conjunto, como uma única obra. No repositório digital, uma obra também pode conter mais de um formato, e o suporte pode ser o HD do servidor, CD-ROM, DVD, fita DAT *etc*. Para simplificar, qualquer desses suportes será tratado apenas como "suporte digital" ou "meio digital".

O trabalho está estruturado da seguinte forma: o Capítulo 2 trata da instituição em que ocorre a implementação do repositório digital de informações, seu histórico, organização e necessidades consideradas para a utilização de um repositório digital de informações; o Capítulo 3 trata desses repositórios digitais, conceituando-os e levantando questões específicas acerca do *software*, do *hardware*, do tipo de profissional necessário ao seu funcionamento e outras questões mais gerais que surgem durante o ciclo de utilização do repositório; o Capítulo 4 apresenta alguns programas disponíveis e analisa a adequabilidade de suas características às necessidades percebidas da instituição; o Capítulo 5 apresenta o *software* escolhido dentre os apresentados e as etapas do processo gerado pela instalação do *software* e implantação do repositório digital até o ponto onde se encontra atualmente. O Capítulo 6 descreve as conclusões deste autor, bem como aponta os projetos a serem desenvolvidos a partir da conclusão deste.

2. A INSTITUIÇÃO

2.1 Histórico

A Igreja Adventista do 7º Dia teve sua origem num grupo de estudiosos da Bíblia que se organizou como igreja em 1863. Esse grupo decidiu estabelecer escolas que, no decorrer dos anos, expandiram sua clientela a todos aqueles que simpatizavam com sua filosofia e seus métodos. No dia 3 de junho de 1872, a educação adventista teve seu início com a abertura da Battle Creek School, Michigan, Estados Unidos, que se destinava a atender os níveis elementar e secundário (EDUCADVENTISTA.ORG, 2006).

No Brasil, em 1896, começou a funcionar em Curitiba, Paraná, o Colégio Internacional, sob a direção de Guilherme Stein Jr. Em 1897, Stein Jr. fundou uma nova escola em Gaspar Alto, SC. A partir daí, o trabalho educacional cresceu e muitas escolas foram agregadas a essa, formando a rede de escolas adventistas (EDUCADVENTISTA.ORG, 2006).

Mundialmente, a educação adventista está presente em 160 países, representada por mais de seis mil instituições da educação infantil à livre docência, totalizando cerca de um milhão de alunos (EDUCADVENTISTA.ORG, 2006).

No Brasil, conta com mais de quinhentas unidades escolares e 128 mil alunos. Além dessas unidades, a organização mantém 12 colégios em regime de internato, da educação básica à superior, e um Centro Universitário em São Paulo, inicialmente conhecido como Instituto

Adventista de Ensino (IAE) e hoje como UNASP Campus São Paulo ou UNASP-SP.

O IAE, inicialmente pequeno, cresceu, e a necessidade de um novo campus tornou-se urgente. Foi adquirido um terreno em Engenheiro Coelho, SP, onde foi construído um segundo campus, denominado Novo IAE (hoje UNASP-EC). Com o objetivo de ainda fundar a Universidade Adventista do Brasil, mais cursos foram abertos, então em dois campi. Educação Artística, Letras, Administração, Nutrição, Matemática, Educação Física, Ciência da Computação, Biologia, Engenharia Civil, Comunicação Social, Contabilidade, Direito, cada uma das faculdades implantadas deu gradativamente corpo ao projeto do Centro Universitário Adventista, nível alcançado no final da década de 1990 (UNASP, 2006).

Recentemente, o UNASP uniu-se ao cinqüentenário Instituto Adventista São Paulo (IASP), em Hortolândia, que passou a ser conhecido como UNASP-HT. Juntos, os três campi somam mais de cinco mil alunos, que cursam do ensino infantil até a faculdade. Somam, também, mais de vinte cursos e oferecem aos alunos a possibilidade de viverem no próprio campus, num conceito de imersão no ensino. Além dos três campi UNASP em São Paulo, há outras instituições adventistas que oferecem cursos superiores em outras partes do Brasil (UNASP, 2006).

Quanto ao uso de um repositório digital, no momento existem iniciativas individuais nessas instituições, mas nada que abrace todo o sistema educacional adventista brasileiro ou mesmo os três campi UNASP. O objetivo inicial é a implantação do repositório no UNASP-SP para a partir de então avaliar as possibilidades de expansão.

2.2 Necessidades gerais identificadas

Com o aumento do número de cursos do UNASP-SP e o incentivo maior à pesquisa, cresceu a produção de conhecimento. Foi estabelecido que todos os alunos deixassem uma cópia impressa de seus trabalhos de conclusão de curso na biblioteca. Esse acervo é fechado, de modo que para acessá-lo o usuário da biblioteca deve ser acompanhado por um funcionário. Como na maioria dos casos há apenas um exemplar de cada trabalho, as obras não podem ser emprestadas, apenas consultadas na biblioteca, já que em caso de extravio a reposição é difícil. Isso gera duas dificuldades: 1) a necessidade de acompanhamento do funcionário tira dele tempo que poderia ser usado para desenvolver outra atividade; 2) muitos usuários não têm tempo suficiente para consultar as obras na biblioteca. Para eles seria mais cômodo e produtivo ter a obra disponível e acessível de qualquer ponto, inclusive de sua residência.

A instituição passou então a solicitar que os alunos entregassem à biblioteca uma cópia do trabalho em disquete ou CD, além da impressa. Uma vez que para cada trabalho impresso haveria outro em meio digital, o digital poderia ser enviado por *e-mail* ou copiado para o usuário que precisasse dele. Mas na realidade isso mostrou-se impraticável devido à grande quantidade de solicitações e ao fato de que o manuseio constante dessas mídias acelera a sua deterioração.

O que motivou este trabalho foi a necessidade imediata de tornar facilmente disponível o texto integral dos trabalhos de conclusão de curso dos alunos de graduação e pós-graduação do UNASP Campus SP. No

entanto, no futuro ele não se limitará a essa facilidade. Os principais objetivos para o repositório digital são:

- Criar uma base dados organizada do conhecimento produzido na instituição;
- Compartilhar esse conteúdo entre instituições, a princípio adventistas, mas sem se restringir a elas;
- Facilitar o processo de construção do acervo da instituição provendo um local único para armazenamento, indexação e recuperação das informações.

2.3 Necessidades específicas quanto ao software / usuário

Considerando-se os objetivos gerais da instituição, foi analisado quais seriam as necessidades específicas quanto ao *software* a ser utilizado. Algumas são imprescindíveis e outras são apenas desejáveis. A Tabela 2 e a Tabela 3, ambas no início do Capítulo 4, sintetizam essas necessidades, comparando-as com as funcionalidades presentes nos aplicativos analisados.

O perfil da maioria dos usuários do repositório de informações não é necessariamente o do pesquisador mais aprofundado, mas basicamente alunos de graduação e especialização, além de funcionários das áreas administrativas da instituição. Portanto, o *software* deve ser fácil de utilizar e não necessitar de nenhuma instalação ou configuração na máquina do usuário final, o que é viável utilizando-se um *software* via *web* (apesar de a instalação de algum *plugin* popular não ser problema).

Na máquina do bibliotecário a questão da instalação e configuração de um *software* não seria problema, já que é um usuário

extremamente interessado no funcionamento do *software*. Além disso, a quantidade de bibliotecários não é tão grande quanto a de usuários finais e o Centro de Informática da instituição oferece suporte técnico para fazer a instalação e configuração necessárias.

A interface do *software* deve ser intuitiva. Caso contrário corre-se o risco de a utilização do repositório ficar aquém do ideal porque o usuário pode não estar disposto a percorrer o processo de instalação, configuração e aprendizado de uma ferramenta mais complexa, mesmo com o benefício de melhores resultados em sua pesquisa.

O *software* deve possibilitar a criação de coleções distintas para facilitar a organização do acervo. A estrutura mínima desejada é a de agrupar as obras em coleções. Por exemplo, as monografias produzidas pelos alunos do curso de psicologia estariam agrupadas na coleção "Monografias do Curso de Psicologia". Seria interessante também que houvesse a possibilidade de hierarquia entre as coleções. Dando continuidade ao exemplo anterior, neste caso poderia ser criada a coleção "Curso de Psicologia" e subordinada a ela as coleções "Monografias" e "Materiais das Disciplinas". Subordinada a esta também poderia ser criada a coleção "Psicologia do comportamento", contendo os materiais dessa disciplina. Seria interessante, também, que o *software* apresentasse comportamentos diferentes para coleções diferentes, pelo menos quanto a texto e imagens. Por exemplo, se uma coleção é composta por texto, o resultado da consulta deveria exibir o título, autor e os arquivos disponíveis. Se fosse uma coleção de imagens, exibir uma lista de miniaturas, títulos e responsáveis pela imagem (fotógrafo, profissional de artes gráficas *etc*.).

Apesar de a estrutura de coleções ser importante, muitas vezes o usuário do repositório prefere navegar por listas e escolher visualizar apenas as obras de determinado autor, por exemplo. Nesse caso, a estrutura hierárquica seria desfavorável. Sendo assim, o *software* deve dar a opção de navegar pelo menos por listagens de autor, título e ano ou data. A opção de navegar por ano é especialmente interessante para a busca pelos trabalhos de conclusão de curso e monografias, onde essa informação é muito relevante; mas não se restringe a esses tipos de material. Pode aplicar-se também a outras obras que podem ficar obsoletas com o tempo, o que é notável, por exemplo, em obras das áreas que envolvem tecnologia. O responsável pela coleção poderia verificar as obras de certo período de tempo para decidir se devem ou não ser retiradas do acervo.

No que diz respeito à recuperação da informação, além da organização por coleções e da possibilidade de navegar por listas diversas, o *software* deve oferecer opções de busca. O mínimo desejável é a busca por autor, título, assunto e palavras-chave. No entanto, quanto maior a possibilidade de refinamento melhor. O ideal seria a opção de busca também pelo texto integral da obra (caso seja texto) e ainda combinações que incluam todos esses campos.

Para o administrador de rede, a facilidade de instalação e de integração com os aplicativos já instalados no servidor é interessante. Assim, se dois aplicativos tiverem funcionalidades semelhantes e o primeiro deles trabalhar com o banco de dados já em funcionamento mas o segundo não, o primeiro terá maior preferência. O mesmo vale para outros requisitos como servidor de internet e linguagens de programação.

É importante ressaltar que é indispensável que todos esses programas sejam livres, já que o local escolhido para sua instalação foi o laboratório de informática, que é baseado em *software* livre, além de não haver interesse por parte da instituição em investir em *software* proprietário, caso o *software* livre atenda às necessidades identificadas.

Muitos *sites* e portais na internet oferecem personalização para o usuário. Geralmente é necessário fazer um pequeno cadastro incluindo o endereço de *e-mail* e a partir de então é possível contar com serviços como alertas personalizados e "carrinhos de compras", dependendo do tipo de produto oferecido pelo *site* em questão. Seria interessante que o *software* do repositório digital oferecesse esse tipo de serviço. Usuários cadastrados teriam acesso a serviços como salvar os resultados de suas buscas, receber por *e-mail* alertas quando uma obra de determinado assunto fosse adicionada ao acervo ou quando determinada obra estiver para ser excluída do acervo. Usuários em geral, mesmo os não cadastrados, poderiam ver na página inicial quais foram as últimas obras acrescentadas ao acervo. Também seria interessante a exibição de estatísticas de utilização para todos os usuários (obras mais visualizadas, obras mais baixadas *etc*.) e outras específicas para o administrador (quais coleções ocupam mais espaço na rede, quais recebem mais acessos *etc*.).

Ainda na questão do cadastro de usuários e personalização de acesso, é importante definir quem pode submeter obras ao repositório e em que nível (autor, revisor, bibliotecário *etc*.). Seria interessante que o próprio responsável pela obra fizesse a submissão. No entanto, para garantir a relevância da obra e a qualidade dos seus descritores (autor, assunto, título, local e data de edição *etc*.), conhecidos como metadados e

que serão discutidos no próximo capítulo, pode ser necessária a participação de pelo menos mais uma pessoa no processo. Por exemplo, um aluno pode submeter um trabalho ao repositório, o seu orientador aprovar ou não a incorporação desse trabalho ao acervo e, se aprovado, o bibliotecário acrescentar os metadados ao registro do trabalho. Se o trabalho não for aprovado, é excluído da base de dados, sendo que o autor deve ser informado antes.

Apesar de não ser objetivo desta monografia, a personalização do acesso abre outras possibilidades de uso para o repositório digital que é interessante citar. O acesso a determinadas coleções pode ser restrito, podem haver coleções públicas e outras disponíveis apenas para usuários internos à instituição ou a certos setores da instituição. Por exemplo, considerando que a secretaria de uma escola não tenha um sistema de informática, ela poderia manter as notas dos alunos no repositório para consulta mais rápida, mas não é interessante que qualquer pessoa veja esses dados. Outra opção é a de cobrar uma taxa mensal de cada usuário por acesso a determinada coleção. Se o *software* do repositório mantiver *logs* necessários, também é possível cobrar uma taxa por cada obra acessada e repassar um percentual dessa taxa para o autor da obra, de modo que este se sinta encorajado a manter atualizado o seu trabalho.

Um dos elementos mais importantes e controversos de uma obra é o assunto. Dois profissionais da informação podem ter opiniões distintas sobre qual é o assunto de uma obra. O assunto também pode assumir facetas ou nomenclatura diferentes de acordo com a instituição ou ainda pode assumir vários termos significando a mesma coisa. Por exemplo, um pesquisador interessado em doenças de coração pode buscar pelo assunto

"doenças de coração", "doenças cardíacas" ou cardiopatias". Todos os termos são válidos e significam a mesma coisa mas, dependendo da formação do usuário que vai pesquisar, um deles pode ser utilizado preferencialmente ou o usuário pode até mesmo não conhecer algum dos termos. É interessante que haja um vocabulário controlado, ou seja, uma lista de termos que devem ser utilizados e uma lista de termos que não devem ser utilizados. De preferência, os últimos devem remeter aos primeiros caso haja equivalência. Por exemplo, se o pesquisador tentar procurar pelo assunto "Cardiopatias" deve ser informado pelo *software* de que o termo utilizado naquela base específica é "Doenças cardíacas". Termos que remetem a outros termos são chamados de *remissivas* pela área da biblioteconomia e ciências da informação. Apesar de ser desejável que o *software* implemente a utilização de remissivas, isso também pode ser implementado operacionalmente, ou seja, a partir de um acordo entre os administradores do conteúdo do repositório.

Nem sempre os usuários do repositório terão acesso à internet no momento exato em que necessitam. Pensando nisso, um item desejável é que o *software* permita a exportação de coleções ou partes de coleções para utilização local. Por exemplo, na rede adventista existem escolas do nível fundamental e médio que não têm condições de oferecer acesso à Internet para seus alunos. Nesse caso, seria possível exportar os itens de interesse para determinada escola de modo que essa parte do repositório pudesse ser acessada a partir de um CD-ROM. Outra possibilidade é um usuário fazer uma busca e exportar os resultados para consulta local posteriormente.

O *software* deve apresentar características de interoperabilidade, permitindo no mínimo a importação e exportação manual de registros bibliográficos de acordo com o padrão de descrição bibliográfica escolhido. Será discutido mais sobre este assunto no Capítulo 3, ao se falar sobre metadados. O ideal é que seja implementado algum mecanismo automático de compartilhamento de obras e metadados.

Finalmente, é interessante que o *software* possa referenciar a obra através de identificadores persistentes, de forma que, mesmo que a obra mude de local onde está hospedada, as referências a ela continuem válidas. Identificadores persistentes podem ser endereços que referenciam uma obra e não mudam mesmo que o endereço do *site* onde a obra está hospedada seja alterado. Isso permite que um repositório-destino importe com segurança apenas os registros bibliográficos de um repositório-origem e não necessariamente o conteúdo integral da obra, economizando espaço em disco). O repositório-origem pode ter seu endereço alterado e mesmo assim o repositório-destino poderá acessar a obra importada, uma vez que no registro bibliográfico importado consta o identificador persistente relativo àquela obra.

3. Repositórios Digitais de Informação

3.1 Conceitos básicos

As bibliotecas têm servido como local de armazenamento, preservação, organização e divulgação de informações desde a sua criação. Em seus primórdios, a informação armazenada estava contida basicamente em livros e seu acesso era excessivamente controlado. Com o tempo, o suporte da informação tornou-se mais variado e a função primordial da biblioteca passou a ser levar a informação ao usuário que necessita dela.

Conforme a tecnologia evoluiu, o suporte da informação diversificou-se ainda mais e as bibliotecas deixaram de trabalhar apenas com livros para receber vários outros tipos de registros informacionais como fotografias, gravações de áudio e/ou vídeo, programas de computador *etc*. A partir da década de 90 principalmente, com a revolução dos computadores e das tecnologias de redes e telecomunicações começou a ser possível construir coleções totalmente digitais. A idéia de compartilhar informações entre bibliotecas também ganhou forças, inicialmente com o compartilhamento de registros bibliográficos, que são descrições padronizadas sobre as obras presentes numa biblioteca, mas não necessariamente a obra em si. Trocando esses registros, as bibliotecas participantes de programas de cooperação tinham condições de conhecer o acervo das outras sem a necessidade da presença física de um profissional. No entanto, para acessar o conteúdo de uma obra ainda seria preciso estar diante dela.

Com o aumento da capacidade de armazenamento dos computadores, bem como o surgimento e evolução da *World Wide Web* e a maior velocidade das redes, começa-se a pensar em repositórios de informação que permitam o acesso ao conteúdo integral das obras em formato digital. (BARTON, 2005) define repositórios digitais institucionais como *"um banco de dados com um conjunto de serviços para capturar, armazenar, indexar, preservar e distribuir a pesquisa de uma instituição de ensino em formatos digitais"*.

Repositórios digitais de informação são, portanto, coleções de obras que podem estar em diversos formatos, mas em suporte digital, e que estejam disponíveis para acesso através do computador. Podem ser acessados a partir de discos magnéticos ou por uma rede de computadores. Não existe limitação quanto ao conteúdo, podendo ser repositórios de informações institucionais (memorandos, projetos *etc*.), acadêmicas (teses, monografias, resenhas) ou qualquer outro tipo de informação que se deseje.

Falando de repositórios digitais institucionais, mais especificamente de universidades, (LYNCH, 2003) os define da seguinte forma:

> *"...um repositório institucional de universidade é um conjunto de serviços que uma universidade oferece aos membros de sua comunidade visando o gerenciamento e disseminação de materiais digitais criados pela instituição e membros de sua comunidade. É essencialmente um compromisso da instituição quanto à responsabilidade sobre esse material digital, incluindo preservação a longo prazo*

quando apropriado, bem como sua organização e acesso ou distribuição...um repositório institucional maduro e bem realizado conterá o trabalho intelectual da faculdade e estudantes – tanto materiais de pesquisa quanto de ensino – bem como documentos sobre a atividade da instituição em si na forma de registro de eventos e desempenho da vida intelectual da instituição".

Conforme (NOERR, 2003), entre os principais objetivos que levam uma instituição a utilizar um repositório digital de informações podem estar:

- Participar de uma comunidade que produz e compartilha conhecimento;
- Aumentar a percepção de valor da biblioteca diante do restante da instituição ou de outras bibliotecas;
- Promover a instituição diante de seus concorrentes e potenciais clientes;
- Gerar renda, cobrando pelo acesso ao acervo.

Além dos motivos apresentados por (NOERR, 2003), outros identificados por este autor seriam:

- Aumentar a disponibilidade da informação, tanto geograficamente quanto no que diz respeito à facilidade de busca e velocidade de acesso;

- Tornar o repositório digital um complemento do acervo da biblioteca física;
- Diminuir a quantidade de visitas dos usuários à biblioteca, o que pode ser desejável caso a única coisa que motive a visita seja a consulta ao conteúdo que poderia estar disponível via *Web*;
- Liberar espaço físico ocupado por materiais que podem ser disponibilizados digitalmente.

3.2 Metadados

Segundo (NISO, 2004), *"metadados são informações estruturadas que descrevem, explicam, localizam ou senão facilitam a localização, uso ou gerenciamento de um recurso informacional"*. Também são chamados de dados sobre dados ou informações sobre informações. Já (ALVARENGA, 2003), define um metadado como *"dado que descreve a essência, atributos e contexto de emergência de um recurso (documento, fonte, etc.) e caracteriza suas relações, visando ao acesso e ao uso potencial"*.

Em outras palavras, metadados são informações sobre determinada obra que abrangem vários contextos como que tipo de material a compõe, responsabilidade de autoria, onde a obra se localiza, onde foi criada *etc*. Existem esquemas que dão estrutura a essas informações, chamados de esquemas de metadados, conjuntos de metadados ou ainda padrões de metadados. Normalmente, um dos objetivos quando esses padrões são criados é garantir a uniformidade na descrição das obras.

Em bibliotecas convencionais um esquema muito usado é o AACR2 (*Anglo American Cataloguing Rules Edition 2*) (GORMAN, 1983), que define os metadados que vão constar nas fichas catalográficas e, conseqüentemente, nos catálogos. Com a popularização dos computadores, o AACR2 passou a ser usado em conjunto com o formato MARC, que é mais facilmente tratado por computador (LIBRARY OF CONGRESS, 2005). Como cada tipo de obra pode ter um esquema mais adequado à sua descrição específica, surgiram vários padrões mais específicos para determinadas áreas que para outras. Por exemplo, "duração em minutos" seria um metadado interessante para um filme ou uma música, mas não faz sentido para um texto impresso ou uma fotografia. Por outro lado, "autor" é um metadado interessante para praticamente qualquer tipo de obra. Normalmente os esquemas de metadados apresentam informações mais gerais coincidentes como autor e título da obra, por exemplo, enquanto abrangem informações mais aprofundadas sobre o tipo de material a que se destinam. A profundidade da descrição desejada pelo profissional da informação também influencia na escolha de um padrão de metadados; quanto mais detalhada a descrição desejada, mais campos o padrão de metadados deverá conter.

Os metadados facilitam a busca nas bases de dados. A busca utilizando metadados ganha em eficiência e relevância, uma vez que é possível determinar quais campos devem ser pesquisados e que valores procurar em cada campo, o que não poderia ser feito num arquivo de fichas catalográficas ou numa busca *full-text*, por exemplo. Comparando mecanismos de busca da internet, que apresentam indexação automática, com busca por metadados criados por seres humanos, (KNIGHT, 2004)

diz que a utilização de metadados retorna resultados relativamente pequenos mas significativos quanto ao detalhamento.

Para obter funcionalidade semelhante, considerando-se uma biblioteca convencional sem computadores que contenha vários tipos de material, se o material desejado fosse um filme, seria necessária a busca manual nas fichas catalográficas de autor, assunto ou título para, aí sim, procurar o filme especificamente. Poderia ser criado um arquivo de fichas a mais, organizado por tipo de material. O mesmo seria necessário caso a busca tivesse que ser realizada por data de publicação. Tudo isso é custoso e ineficiente. Mesmo se a busca fosse eletrônica, mas simplesmente acessando o conteúdo completo da obra, seria difícil localizar precisamente o que se deseja. E há tipos de materiais que não podem ser encontrados através de uma busca *full-text* como, por exemplo, programas já compilados. A busca eletrônica por meio de metadados acelera o processo e restringe a pesquisa, aumentando a eficiência e relevância dos resultados. Conforme (NISO, 2004), o uso de metadados apresenta como funcionalidades permitir que os recursos sejam encontrados a partir de critérios relevantes, agrupar recursos semelhantes e, conseqüentemente, segregar recursos com características diferentes, além de prover informações sobre a localização dos recursos. Por recursos, pode-se entender obras em quaisquer formatos, digitais ou não.

Segundo (COLE, 2002), os seguintes princípios auxiliam na construção de metadados de qualidade:

- Devem ser apropriados ao material armazenado, usuário do acervo e uso pretendido para o acervo;
- Devem suportar interoperabilidade;

- Devem ser confiáveis;
- Devem suportar gerenciamento a longo prazo dos objetos digitais.

Como (COLE, 2002) cita, a interoperabilidade também é uma questão que depende de um padrão de metadados. Se duas bibliotecas têm a mesma obra, é possível e desejável que apenas uma delas passe pelo processo descrever a obra. Desde que o sistema de informática de ambas utilize o mesmo padrão de metadados – ou ofereça a possibilidade de tradução entre padrões – a descrição feita em uma biblioteca pode ser importada para a base de dados da outra, evitando o retrabalho.

Num ambiente digital, caso entre os elementos do padrão de metadados adotado conste a localização eletrônica da obra, basta que se importem os metadados, não necessariamente a obra em si. Assim, o pesquisador recebe como resultado de sua busca local um endereço para uma obra que se encontra armazenada em uma base de dados remota. A base local apresenta um conteúdo mais rico sem necessariamente hospedar a obra. Torna-se possível inclusive ter um repositório digital sem nenhuma obra de fato, ou seja, uma compilação de registros bibliográficos que sejam pertinentes à área de conhecimento abrangida pelo repositório (que pode ser especializado em informações sobre engenharia genética, por exemplo) e cujos metadados remetam à obra em si, armazenada em outro local. Presta-se o serviço desejado – fornecer informação relevante na área de atuação escolhida - mas com economia de recursos.

Os esquemas de metadados contém elementos que recebem nomes de acordo com o contexto (autor para livro e compositor para uma música, por exemplo) e valores (o nome do autor/compositor propriamente dito). Assim, os *elementos* dos metadados recebem *valores*. Normalmente existem regras de sintaxe para os elementos e pode ser interessante especificar regras para o conteúdo dos valores a fim de garantir a interoperabilidade.

Exemplificando a questão do conteúdo dos valores, se um livro apresenta dois títulos na capa, qual dos dois deverá ser considerado realmente o título principal do livro? Se não houver nenhuma regra, duas pessoas responsáveis pelo preenchimento dos valores dos elementos de metadados podem escolher títulos distintos para uma mesma obra. Ou ainda, uma mesma pessoa pode escolher títulos distintos, caso trabalhe a mesma obra em duas ocasiões diferentes. O mesmo vale para o assunto da obra, que é ainda mais passível de interpretações distintas. O AACR2, apesar de não se referir especificamente a metadados para objetos digitais, trata de questões de padronização desses valores, podendo ser uma ferramenta interessante dependendo do ambiente em que está inserido o repositório. Quanto à sintaxe, os esquemas atuais voltados às obras digitais geralmente utilizam a linguagem SGML (*Standard Generalized Mark-up Language*) (W3C, 2004) ou XML (*Extensible Mark-up Language*) (W3C, 2006), desenvolvidas pelo *World Wide Web Consortium* (W3C) e amplamente usadas para a troca de informações estruturadas.

No contexto dos repositórios digitais, existem duas possibilidades de armazenar metadados: armazenar os metadados dentro dos próprios

arquivos da obra ou ligar os metadados à obra correspondente. Páginas HTML, por exemplo, são candidatas perfeitas a armazenar os metadados dentro da própria obra. Alguns formatos de arquivo também possuem campos para metadados. Em imagens JPEG, por exemplo, há informações EXIFF; em arquivos de áudio no formato MP3, existem as *tags* ID3. Por outro lado, não é todo tipo de arquivo que permite armazenar metadados internamente, ou as possibilidades de descrição oferecidas são limitadas. Organizar os metadados externamente à obra proporciona maior flexibilidade ao mesmo tempo em que possibilita uma interface mais genérica para armazenamento e recuperação desses metadados.

A seguir são discutidos conceitos quanto à questão dos metadados e interoperabilidade entre sistemas e apresentadas tendências e padrões mais recorrentes atualmente no universo dos repositórios digitais dentro do contexto deste trabalho.

3.2.1 *Open Archives Initiative* (OAI)

Segundo (OPEN ARCHIVES FORUM, 2003), a *Open Archives Initiative* (Iniciativa dos Arquivos Abertos ou simplesmente OAI) *"desenvolve e promove padrões de interoperabilidade que visam facilitar a disseminação eficiente de conteúdo"*. O objetivo é prover a interoperabilidade entre repositórios através do compartilhamento, disseminação e armazenamento de metadados e materiais digitais.

No contexto da OAI existem dois tipos de funções não mutuamente exclusivas que um repositório pode assumir: provedores de dados e provedores de serviços. Os primeiros disponibilizam seus metadados para os últimos. Estes coletam metadados de várias fontes e o

serviço oferecido é baseado nessas informações coletadas. Os metadados coletados podem estar em qualquer formato, apesar de ser necessário o uso de elementos básicos do padrão Dublin Core, discutido mais adiante, para garantir o mínimo de interoperabilidade (OPEN ARCHIVES FORUM, 2003).

Como todo tipo de comunicação requer um protocolo apropriado, o protocolo desenvolvido pela OAI é o *OAI-Protocol for Metadata Harvesting* (Protocolo OAI para coleta de metadados). Ele funciona baseado nos padrões HTTP e XML. Não define como será feita a busca dos dados, apenas como os dados deverão ficar em um só lugar.

As requisições OAI-PMH que podem ser feitas a um repositório utilizam seis verbos, apresentadas na Tabela 1 (PRASAD, 2005):

Tabela 1 – Lista de verbos possíveis numa requisição OAI-PMH

Verbo	**Retorno**
Identify	Descrição do repositório
ListMetadataFormats	Formato de metadados suportado pelo repositório
ListSets	Agrupamentos (de metadados) definidos pelo repositório
ListIdentifiers	Identificadores do item
ListRecords	Apresenta registros do repositório
GetRecords	Adquire registros do repositório

O repositório pode ter a opção de responder automaticamente a requisições OAI-PMH, ou seja, se o repositório implementa o protocolo OAI-PMH, outros repositórios podem fazer requisições em seu banco de metadados (OAI, 2004).

O interessante do uso de HTTP e XML é que os dados dos repositórios, se transmitidos e armazenados nesses formatos, ficarão visíveis a mecanismos de busca como o Google. Caso contrário, para encontrar uma das obras que constam no repositório, seria necessário que o mecanismo de busca utilizado pelo usuário traduzisse os termos pesquisados para a sintaxe de busca do repositório em questão. Se isso não acontece, o material armazenado no repositório fica escondido em seu interior. Como existem milhares de repositórios com sintaxes diferentes, há muito material "invisível" externamente a esses repositórios, o que é conhecido como *deep web* (*web* profunda ou *web* oculta) e, segundo estudos, é várias vezes maior que a parte mais superficial da *web* (BERGMAN, 2001). Portanto, quando se fala em disseminação neste contexto, refere-se à possibilidade de o documento ser "extraído" da *deep web* e ficar visível através de mecanismos de busca mais gerais. Atualmente existe inclusive um módulo para tornar o Apache capaz de responder a solicitações OAI-PMH (NELSON, 2004) .

3.2.2 MARC

O formato MARC foi desenvolvido pela *Library of Congress* (Biblioteca do Congresso Norte Americano) que é a fonte mais confiável para catalogação dos Estados Unidos, portanto uma das autoridades mais reconhecidas no mundo nessa área. Considerou-se que os computadores não podem ler diretamente as fichas catalográficas que eram feitas utilizando-se as normas AACR2. Assim, era necessário um esquema que fosse legível por computador (LIBRARY OF CONGRESS HELP DESK, 2003).

A estrutura do formato MARC é descrita a seguir, conforme indicado por (LIBRARY OF CONGRESS HELP DESK, 2003): os registros bibliográficos são compostos por campos como autor, assunto *etc*. A estrutura do formato MARC é basicamente a seguinte: cada campo é associado a um número de 3 dígitos, chamado de *tag MARC*. Assim, cada *tag* identifica um campo ou uma área. Dessa forma, o computador pode identificar, a partir das *tags*, a que campo/área pertence o valor que vem a seguir.

Os 3 dígitos da *tag* são seguidos por 2 dígitos decimais que podem ser usados como indicadores e que são escritos juntos, mas são independentes, ou seja, são 2 dígitos seguidos e não um número com 2 dígitos. Por exemplo, a *tag* de título acompanhada do indicador apropriado pode significar que o computador deve ignorar determinada parte do título no momento de ordenar os registros. Assim, poderia ser indicado que devem ser ignoradas as três primeiras letras do título "*The Matrix"*, ou seja, a parte interessante desse título, para fins de ordenação, seria apenas *Matrix*. Alguns campos podem não ter indicadores, ou apenas um dos indicadores definidos. Se o indicador é indefinido, utiliza-se o caracter # significando espaço em branco.

Um campo pode ter subcampos. Uma obra pode, por exemplo, ter título, subtítulo e um título alternativo – situações previstas no AACR2 - ou o campo que indica a responsabilidade intelectual da obra pode apresentar várias recorrências como autor, tradutor, editor *etc*. Normalmente os subcampos são indicados por um delimitador, que é uma combinação de caracteres, como por exemplo $a ou $b ou $c).

Aplicando os 3 conceitos (campos, indicadores e subcampos), o registro do campo *título* ou da *área de título* de uma obra, que é a nomenclatura utilizada pelo AACR2, poderia ser representado em formato MARC como segue:

245 14 $a Nos caminhos da vida
$b a história de José da Silva.

Onde 245 é a *tag* MARC para toda a área de título (que pode incluir título principal, subtítulo, título original em outro idioma, título alternativo *etc*.), o número 1 é um indicador que significa que deve haver uma ficha no catálogo (convencional) de títulos para esta obra, 4 é um indicador que diz ao computador para ignorar as quatro primeiras posições (no caso, o termo *Nos* mais um espaço em branco) ao ordenar os registros, $a é o título principal e $b é o subtítulo.

Apesar de não ter sido desenvolvido especificamente para objetos digitais, o formato MARC, mesmo sendo um padrão já antigo, é robusto e flexível o suficiente para ser adaptado, além de bem estabelecido (BRANTON, 2004)

3.2.3 METS

Em uma biblioteca convencional, não manter metadados estruturais sobre uma obra não é necessariamente um problema, uma vez que, caso se percam os metadados, a obra não perderá seu conjunto. Já obras digitais, sem metadados estruturais, correm o risco de se "desintegrarem", ou seja, não haver ligação entre as partes que a

compõem (diversos artigos de uma revista, por exemplo). Outro caso: se uma obra foi digitalizada, a ausência de metadados técnicos sobre o processo de digitalização pode fazer perder dados importantes sobre a obra original como. Por exemplo, se um livro foi produzido por uma prensa de Guttenberg e posteriormente digitalizado, a omissão dessa informação nos metadados da obra digital faz com que, para o pesquisador, ela perca muito de seu valor (LIBRARY OF CONGRESS, 2005).

Essas questões foram premissas que levaram a *Library of Congress* a desenvolver o *METS - Metadata Encoding and Transmission Standard* (Padrão de codificação e transmissão de metadados), que inclui elementos *"administrativos e estruturais para trabalhos textuais e baseados em imagens"* (LIBRARY OF CONGRESS, 2005). Ainda segundo (Library of Congress, 2005), "*o METS provê um formato XML para codificar metadados necessários tanto para gestão de objetos de bibliotecas digitais num repositório quanto para a troca desses objetos entre repositórios (ou entre repositórios e seus utilizadores)"*. Segundo (GARTNER, 2002) *"o METS é um padrão emergente projetado para codificar todos os metadados necessários para uma descrição completa de objetos digitais em uma biblioteca digital"*.

Um documento METS é uma aplicação do padrão XML, que é não-proprietário e independente de um *software* específico (NISO, 2004). Descreve um objeto de uma biblioteca digital e consiste em 7 seções principais, resumidas a seguir conforme (LIBRARY OF CONGRESS, 2005) e (NISO, 2004):

1. Cabeçalho: contém metadados sobre o documento METS em si, como autor, editor *etc*.;
2. Metadados descritivos: podem ser internos ao documento ou apontar para registros externos, inclusive em outro formato que não o próprio METS (MARC, Dublin Core *etc*.);
3. Metadados administrativos: descreve como os arquivos que compõem a obra foram criados e armazenados, direitos de *copyright*, metadados sobre o objeto original de onde deriva o que está sendo tratado *etc*. Também podem ser internos ou externos ao documento METS;
4. Seção de arquivos: Lista os arquivos que formam o objeto digital em si (a substância da obra);
5. Mapa estrutural: apresenta a hierarquia entre os componentes do objeto digital e faz a ligação desses componentes aos arquivos e metadados de cada componente;
6. Ligações estruturais: registra a existência de *hiperlinks* entre nós da hierarquia do mapa estrutural;
7. Comportamento: pode ser usada para associar comportamentos executáveis a conteúdos do objeto METS. Permite que se registrem informações sobre como os componentes do objeto digital serão renderizados para o usuário, incluindo que aplicativos devem ser usados ou parâmetros específicos necessários ao renderizar um arquivo.

3.2.4 Dublin Core

Tradicionalmente, a criação de metadados compartilháveis por computador era feita através de registros MARC. No entanto, o uso do MARC exige uma quantidade de treinamento que muitas vezes inviabiliza sua utilização para o público não especializado. Nesse contexto emergiu o padrão Dublin Core, que é forte candidato a se tornar padrão para objetos digitais cuja descrição não necessite de um detalhamento extremamente aprofundado (NISO, 2004).

O formato Dublin Core é um esquema de metadados desenvolvido pela *Dublin Core Metadata Initiative* (DCMI) com o objetivo de tratar informação digital. Segundo (HANSEN, 1999), o formato Dublin Core foi criado visando descrever um recurso eletrônico (local, formato *etc*), tornar a descrição simples e possibilitar a indexação, facilitar a pesquisa de recursos e o acesso a esses recursos. Assim, entre as razões para se adotar o formato Dublin Core, estão:

- facilidade de criação;
- simples de indexar;
- indexação mais precisa do que busca *full-text*;
- interoperabilidade.

A simplicidade do Dublin Core decorre do fato de que ele é composto por apenas 15 elementos, todos opcionais e passíveis de repetição, que podem ser codificados em XML (NISO, 2004). Cada elemento contém um valor que é usado para descrever o recurso

eletrônico em determinado aspecto. Esses 15 elementos são (DCMI, 2004):

- *TITLE*: é o nome que o criador ou publicador dá para a obra;
- *AUTHOR* ou *CREATOR*: a pessoa ou organização responsável pelo conteúdo intelectual da obra;
- *SUBJECT* ou *KEYWORDS*: palavras-chave, assuntos conforme um vocabulário controlado ou não, descritores de classificação *etc*.;
- *DESCRIPTION:* uma descrição textual da obra;
- *PUBLISHER:* Entidade responsável por disponibilizar a obra como uma editora, universidade *etc*.;
- *OTHER CONTRIBUTORS:* Outros responsáveis pelo conteúdo intelectual da obra, diferente de *AUTHOR* ou *CREATOR*.
- *DATE:* Data em que a obra se tornou disponível;
- *RESOURCE TYPE*: tipo de obra. Por exemplo: tese, apostila, dicionário, filme *etc*. É interessante especificar exatamente quais dessas opções estarão disponíveis ao catalogador;
- *FORMAT:* a representação dos dados da obra ou, em outras palavras, o tipo de arquivo. Podem ser usados os tipos de arquivo MIME, conforme definido pelo padrão RFC2046 (RFC2046, 1996);
- *RESOURCE IDENTIFIER:* texto ou número usado para identificar inequivocamente a obra. Exemplo são as URLs e URNs;

- *SOURCE*: o trabalho original, eletrônico, impresso, ou gravação de áudio em vinil, por exemplo, de onde a obra é derivada;
- *LANGUAGE*: Língua em que a obra foi produzida;
- *RELATION*: relacionamento da obra que está sendo descrita com outras obras;
- *COVERAGE*: descreve características espaciais e temporais da obra, quando aplicável;
- *RIGHTS MANAGEMENT*: a idéia é ligar este elemento a uma URL ou URI que apresente os termos de direitos de uso da obra.

Esse esquema que apresenta apenas 15 elementos normalmente é chamado de *unqualified* Dublin Core (Dublin Core não-qualificado). Devido à necessidade de descrições mais completas para os recursos digitais, foi criado um conjunto de qualificadores, de uso opcional mas interessante. Os qualificadores servem para informar ao usuário como enxergar ou interpretar um valor (conteúdo) num elemento (campo do metadado). Também são usados para aprimorar a semântica do conteúdo de um elemento. Os qualificadores podem ser usados para refinar o significado de valores dentro dos elementos, mas não para estender qualquer elemento (HANSEN, 1999). Por exemplo, o elemento *DATE* pode ser usado com o qualificador *SCHEME*, que define o formato em que a data está sendo escrita, podendo ser, por exemplo, o padrão ISO-8601, que descreve formatos de data e hora (ver http://www.w3.org/TR/NOTE-datetime).

O elemento *COVERAGE* é outro bom candidato ao uso de qualificadores. O qualificador *SCHEME* poderia especificar que o valor do elemento está em graus, em minutos, em metros ou mesmo que cita um período histórico como a Idade Média, por exemplo (AD HOC WORKING GROUP, 1997).

Descrever o processo criação dos metadados vai além do escopo deste trabalho. No entanto, vale ressaltar novamente que metadados de qualidade garantem ao usuário utilizador do repositório maior precisão na busca e maior relevância dos resultados. Além disso, apesar de haver vários padrões prontos, a instituição pode escolher como compor o seu próprio esquema de metadados, seja utilizando partes de um padrão já definido, seja combinado ou utilizando simultaneamente dois ou mais padrões (NSDL, 2005).

3.3 Necessidades de *software*

Uma vez que o objetivo da instituição é oferecer acesso via *Web*, os programas que suportam o repositório digital de informações vão necessitar do apoio de um servidor de páginas *web*. Além disso, os objetos digitais devem ficar organizados no servidor, o que, devido a quantidade de objetos e informações associadas a eles, pressupõe a necessidade de um servidor de banco de dados. O contato com o usuário por mensagens de *e-mail* automáticas também necessita de MTA (*mail transport agent)* instalado e funcionando adequadamente.

O esquema de metadados implementado também deve ser compatível com padrões abertos, o que não é problema já que tradicionalmente os esquemas de metadados têm como objetivo a

padronização, sendo disponibilizados imediatamente após o seu desenvolvimento. Em se tratando de repositórios digitais, os metadados costumam ser representados através de linguagens de marcação como XML ou SGML, por exemplo. Mesmo o formato MARC 21 é estruturado em *tags*. Para garantir que os dados digitados ou gerados pelo repositório estejam sintaticamente corretos, aplicativos que realizam a checagem da sintaxe da linguagem de marcação – os *parsers* – deverão estar presentes no ambiente do repositório.

A indexação dos documentos é normalmente feita a partir do texto plano. Por isso, um outra necessidade de *software* são os conversores dos diversos formatos de arquivo para o formato texto. Existem conversores para os vários formatos como DOC, XLS, PPT, PDF, HTML *etc*. O *software* do repositório digital deverá trabalhar em conjunto com, ou implementar internamente, conversores de documentos para o formato texto, a fim de indexar o texto integral das obras.

Dependendo de como o *software* tenha sido escrito, pode ser necessária a instalação de linguagens de programação ou interpretadores de comando. É possível também que o *software* tenha sido disponibilizado apenas como código-fonte, sendo necessária a presença de compiladores. Em ambiente Linux, isso não costuma ser problema.

3.4 Necessidades de hardware

O *hardware* mais óbvio necessário é o servidor onde estará hospedado o repositório. Sua configuração vai depender da performance do *software* escolhido mais a quantidade de acessos simultâneos que o administrador do repositório prevê.

São os objetivos da instituição para o repositório que definirão os demais itens de *hardware*. Se a estratégia for digitalizar conteúdo, ou seja, não incluir na base de dados apenas material que já tenha nascido digital, serão necessários também outros equipamentos e possivelmente outros computadores. Dependendo da quantidade, da origem e da qualidade final desejada para o material a ser digitalizado, o porte e o preço desses equipamentos pode variar. Por exemplo, pode ser desejável digitalizar muitas páginas, mas com qualidade pequena. Ou digitalizar documentos que estejam em formato de livro e que sejam raros, sendo necessário muito cuidado em seu manuseio e qualidade de digitalização para preservar os detalhes. Existem também informações preservadas em discos de vinil, vídeos em VHS, microfilmes *etc.* A instituição deve considerar também a possibilidade de terceirizar esse tipo de trabalho.

3.5 Necessidades de pessoal

Quanto ao pessoal, um repositório digital demanda já antes de estar em funcionamento, equipes de estudo a fim de definir, entre outras coisas, sua política de construção do acervo, política de acessos, manutenção técnica, manutenção do conteúdo em si e participantes do processo de submissão. Nessas equipes também é interessante incluir os usuários finais, para obter uma perspectiva do funcionamento do repositório a partir de seu público-alvo.

Num repositório em funcionamento, a necessidade de pessoal varia de acordo com as políticas pré-estabelecidas, mas tipicamente os seguintes profissionais seriam necessários:

- Administrador de informática: responsável pela manutenção da rede, da instalação e configuração do *software* e por garantir a disponibilidade do repositório;
- Bibliotecário: responsável pela qualidade dos metadados, pela divulgação do repositório em si e de seu conteúdo, por disseminar a informação de maneira geral, e por manter contato com instituições que podem ser parceiras no processo de troca de itens de acervo;
- Autores: responsáveis por alimentar de fato o repositório com conteúdo. Apesar de não ser necessariamente um profissional, obviamente desempenha um papel fundamental no contexto do repositório;
- Revisores: responsáveis por avaliar o trabalho dos autores, sugerir correções ou mesmo negar a inclusão de uma obra no repositório. Devem trabalhar em constante contato com os bibliotecários a fim de, juntos, seguirem a política de construção de acervo. Por exemplo, no caso de trabalhos de conclusão de curso, os revisores poderiam ser os orientadores.

Além disso, caso a estratégia adotada pela instituição seja a digitalização, são necessárias pessoas que trabalhem suas áreas específicas: revisores para conferir se o reconhecimento ótico de caracteres aconteceu de forma correta (no caso de texto), profissionais de audiovisual (para obras em áudio e vídeo), especialistas em aplicativos 3D (digitalização de artefatos) e profissionais da área gráfica (digitalização/correção de imagens).

4. APLICATIVOS ANALISADOS

Este capítulo trata da análise das ferramentas de *software* livre disponíveis para suportar a implementação de um repositório digital. Como todos os aplicativos analisados são livres e possuem interface *web*, estes dois itens não serão comentados repetidamente, apesar de possivelmente haver comentários curtos sobre o tipo de licença e como é o *layout* da interface.

Os requisitos indispensáveis serão analisados um a um. Quanto aos apenas desejáveis, serão comentados durante a descrição de cada ferramenta, bem como quaisquer outras particularidades interessantes da ferramenta. Os pré-requisitos de *software* específicos ao funcionamento do repositório (servidor *Web*, banco de dados *etc*.) são apresentados em uma tabela no Apêndice A e por isso não serão citados durante a análise, a não ser que haja alguma peculiaridade.

A descrição dos aplicativos será realizada, portanto, levando-se em conta os seguintes aspectos (não necessariamente nessa ordem):

- Histórico do *software*;
- Organização funcional;
- Formatos de arquivo suportados e de que forma;
- Opções de navegação pelas coleções;
- Opções de busca;
- Processo de submissão dos arquivos;
    - Submissão do arquivo em si;
    - Inclusão dos metadados;
- Interoperabilidade;

- Protocolos de comunicação suportados;
- Exportação dos metadados;

- Interação com o usuário;
- Documentação.

4.1 Comparação dos aplicativos

A Tabela 2 resume a lista de requisitos apresentados até aqui e define o que é obrigatório - o *software* que não atender nesse sentido será automaticamente eliminado da lista de opções;

Tabela 2 – Comparação dos requisitos obrigatórios da instituição com os recursos dos aplicativos analisados.

Necessário	Dspace	CDSWare	Nou-rau	Greenstone
Software livre	✔	✔	✔	✔
Interface *Web*	✔	✔	✔	✔
Permitir inclusão de mais de um formato de arquivo por obra	✔	✔	x	✔
Fácil utilização	✔	✔	✔	✔
Criação de coleções distintas	✔	✔	✔	✔
Possibilidades de navegar pelos campos autor, título e ano ou data	✔	✔	✔	✔
Busca por autor, título, assunto e palavras-chave	✔	✔	✔	✔
Submissão pelo próprio autor	✔	✔	✔	x
Inclusão de metadados obedecendo algum padrão internacional	✔	✔	✔	✔
Importação e exportação de obras/metadados	✔	✔	✔	✔

✔ =Apresenta funcionalidade; x = não apresenta funcionalidade.

Tabela 2 - (continuação)

Necessário	DSpace	CDSWare	Nou-rau	Greenstone
Interação com usuário (notificações por *e-mail*, informações na *home page*)	✔	✔	✔	x
Quantidade satisfatória de documentação	✔	✔	x	✔

✔ =Apresenta funcionalidade; x = não apresenta funcionalidade.

A Tabela 3 resume a lista do que é apenas desejável – pesa na escolha do *software* mas não é absolutamente necessário.

Tabela 3 – Comparação dos itens desejáveis para um repositório digital institucional com os recursos dos aplicativos analisados.

Desejável	Dspace	CDSWare	Nou-rau	Greenstone
Não necessitar de plugins na máquina do usuário final	✔	✔	✔	✔
Comportamentos diferentes para coleções diferentes	✔	✔	x	✔
Interface agradável e personalizável	✔	✔	✔	✔
Hierarquia entre coleções	✔	✔	✔	✔
Possibilidades de navegar por vários ou todos os campos (além de título, autor e data)	x	✔	x	✔
Busca complexa combinando campos específicos de metadados e texto-integral	✔	✔	x	x
Workflow básico	✔	✔	✔	x
Interoperabilidade automática com outros sistemas (mesma base ou bases diferentes	✔	✔	x	✔
Permite configurar utilização de vocabulários controlados	✔	x	x	✔
Exportar coleções ou partes de coleções para consulta local	✔	x	x	✔
Utilização de protocolo automatizado de troca de registros	✔	✔	✔	✔
Identificadores persistentes	✔	x	x	x

Tabela 3 – (Continuação)

Desejável	Dspace	CDSWare	Nou-rau	Greenstone
Estatísticas de utilização	✔	✔	✔	x
Interação estilo portal: usuários cadastrados têm acesso personalizado	✔	✔	✔	x

✔ =Apresenta funcionalidade; x = não apresenta funcionalidade.

Os seguintes aplicativos foram descartados antes de uma análise mais detalhada porque a própria documentação já demonstrou claramente que não atenderiam aos requisitos indispensáveis. Apesar disso, são listados a título de informação, acompanhados do motivo pelo qual foram descartados:

- GNU/E-Prints (http://www.eprints.org): o objetivo primário é a literatura cinzenta (teses, monografias, relatórios *etc*.). Permite a inclusão de outros tipos de objetos digitais, mas o foco real é a literatura científica;
- Open Journal Systems (http://pkp.sfu.ca/ojs/): é voltado principalmente à publicações periódicas de modo geral, como revistas por exemplo;
- Fedora (http://www.fedora.info): por causa do altíssimo grau de detalhamento oferecido para os metadados, sua utilização é excessivamente complexa para o tipo de usuário-alvo do repositório. É extremamente flexível para suportar os mais variados tipos de objetos digitais. Até o momento, porém, o *software* não tem uma interface própria com o usuário, mas pode ser utilizado como suporte para outras aplicações.

4.2 DSpace

O nome completo do *software* é *DSpace Institutional Digital Repository System* (DSpace – que significa *Digital Space*, ou espaço digital – Sistema de Repositório Institucional Digital). O DSpace é uma base de dados desenvolvida pelo *Massachussets Institute of Technology* (MIT) em parceria com a *Hewllet-Packard* (HP). Atualmente, é um *software open-source* baseado no modelo de licença GPL.

Segundo os próprios desenvolvedores, o DSpace captura, distribui e preserva produtos digitais de pesquisa (DSPACE SYSTEM DOCUMENTATION, 2005). Permite armazenar, indexar e recuperar artigos, teses, relatórios, documentos técnicos, conferências e outros tipos de material digital em vários formatos (inclusive áudio e vídeo). Os documentos podem ser acessados integralmente a partir da interface *web*. Apesar de ter sido desenvolvido visando empresas, seu maior uso tem se dado em universidades, para compartilhar produção científica.

Em termos de desenvolvimento a longo prazo, o DSpace é um *software* interessante por envolver tanto a parceria MIT/HP quanto uma grande comunidade de desenvolvedores. Além disso, a comunidade que gera acervo também é grande, haja vista o número de instituições (não apenas universidades) nacionais e internacionais que aderiram ao seu uso, e a grande variedade e diversidade de conteúdo que pode ser armazenado na base. Atualmente, além do MIT, mais de uma centena de instituições já utilizam o DSpace, inclusive brasileiras (DSPACE INSTANCES, 2006).

A idéia que motivou o desenvolvimento do DSpace baseia-se nas seguintes premissas:

- Muito do material que nasceu digital já está perdido;
- A maior parte do material digital corre riscos;
- É melhor preservar digitalmente que perder completamente;
- É necessário capturar tanta informação quanto possível para suportar a preservação funcional;
- Relação custo/benefício favorável.

A base implementa padrões amplamente aceitos internacionalmente, como o Dublin Core para metadados e o protocolo OAI-PMH para compartilhamento de registros. Atualmente encontra-se disponível um módulo para exportação no formato METS. Dessa forma os acervos armazenados no DSpace podem ser exportados e compartilhados tanto entre repositórios DSpace quanto outros repositórios compatíveis com o protocolo OAI-PMH.

O *software* está atualmente hospedado no endereço http://www.dspace.org. A estrutura do DSpace é a seguinte:

- Comunidades: São as pessoas que pesquisam ou submetem conteúdo à base;
- Coleções: Agrupamentos de obras semelhantes ou relacionadas, pertencentes a comunidades;
- Itens: são as obras, ou objetos digitais, em si; é o que o usuário normalmente está buscando. Apresentam identificadores persistentes;
- *Bitstreams*: Arquivos de computador que compõem cada item.

Em outras palavras, a base de dados é composta por comunidades. Cada comunidade tem suas coleções. Essas coleções são compostas por itens. Cada item apresenta um ou mais arquivos em formatos variados (PDF, JPG, áudio, vídeo *etc*.), chamados de *bitstreams,* com o conteúdo propriamente dito. Cada item também apresenta metadados como título, autor, data em que o documento foi aceito, palavras chave, resumo, em quais coleções aquele item aparece (um mesmo item pode pertencer a mais de uma coleção de comunidades diferentes), tamanho em *bytes* de cada *bitstream*, entre outros. Esses dados podem ser exibidos de forma completa ou resumida.

O DSpace apresenta também características de *workflow*, ou seja, existe um administrador que pode aceitar ou não a submissão de um documento à base de dados.

Ao se exibir o resultado da busca por determinado item é apresentado o identificador persistente, que é um endereço de internet baseado no sistema *Handle* (http://www.handle.net) que garante que aquele item será sempre encontrado através do endereço eletrônico correspondente. A Figura 4.1 mostra a exibição dos metadados básicos de um item. A primeira informação é o identificador persistente. Logo, ao se citar este item em uma referência bibliográfica, pode-se usar o identificador http://hd1.handle.net/1721.1/3541 com a certeza de que esse endereço será sempre referente à obra com o título *Welfare Implications of User Innovation.*

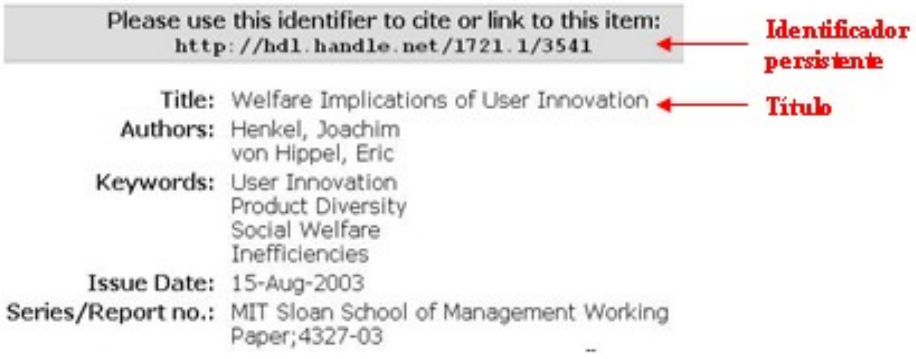

Figura 4.1 - Apresentação básica dos metadados de um item no DSpace

O usuário pode pesquisar diretamente digitando termos ou optar por navegar pelas comunidades e suas respectivas coleções, ou ainda navegar por título, autor ou data. Existe também a opção de se cadastrar para receber por *e-mail* as novidades de cada coleção, sempre que itens forem adicionados ou alterados. O *software* também apresenta áreas livres a qualquer usuário e áreas restritas. Só usuários autorizados podem submeter itens ao acervo. Dependendo do acervo, a submissão de um item pode ser livre ou passar por mediadores, que autorizam ou rejeitam a entrada desse item no acervo.

Quanto à busca, como citado, existe a opção de navegar pelas comunidades/acervos, títulos, autores e data. Existem também a opção de digitar termos para uma busca geral e a opção muito interessante de digitar termos para buscar apenas dentro dos acervos de determinada comunidade. Na Figura 4.2 é apresentada a interface de busca do DSpace.

Figura 4.2 – Interface de busca do DSpace

Segue-se uma explicação das opções da tela apresentada na Figura 4.2:

1. Área para a busca geral em toda a base de dados DSpace. A busca avançada permite que se especifiquem em que campos buscar e combinar essas buscas com os operadores booleanos "*and*", "*or*" ou "*not*". Além disso, permite que se procurem por palavras-chave que, no contexto do DSpace significa busca em texto integral, além dos metadados;
2. Área que permite navegar pelas comunidades e seus acervos; ou por títulos, autores ou data.
3. Nome da comunidade corrente (comunidade na qual o usuário está realizando a sua busca. Neste caso, *Sloan Working Papers*.

4. Área em que o usuário cadastra seu *e-mail* para receber novidades sobre um acervo da comunidade.
5. Área em que os usuários (apenas os autorizados) podem submeter documentos para um acervo da comunidade.
6. Área para a busca específica, somente na comunidade corrente. Caso o usuário opte por navegar nesta área, serão exibidos apenas os itens pertencentes à comunidade especificada.
7. Área em que são exibidas as submissões mais recentes ao acervo.

Nas buscas em que se digitam termos, sejam elas gerais (área 1) ou específicas por comunidade (área 6), a pesquisa é realizada da seguinte forma:

- Algumas palavras são ignoradas na busca. Por ser uma base em inglês, algumas delas são: "*a*", "*and*", "*are*", "*as*", "*to*", "*was*" *etc*;
- O asterisco é usado para truncar. Assim, a busca test* deve retornar testando, testes, testamento *etc*;
- As palavras têm seu final expandido com os finais mais comuns para recuperar mais resultados (plural, verbo no passado *etc*.);
- Frases devem ser delimitadas por aspas;
- O símbolo (+) adicionado à frente de uma palavra indica que ela deve obrigatoriamente aparecer no resultado da busca.

Assim, a busca *+bibliotecas digitais* indica que *digitais* é opcional, mas *bibliotecas* deve aparecer no resultado;

- O símbolo (-) é usado para negação;
- Os operadores booleanos (*AND*, *OR* ou *NOT*) são suportados, mas para isso devem ser escritos em caixa alta;
- Podem-se usar parênteses para combinar grupos estratégias de busca. Por exemplo: (publicidade OR *marketing*) AND (biblioteca OR "centro de informações").

A interface com o usuário é simples e funcional. Como pode ser observado nas figuras apresentadas, o acesso a cada função é facilmente visível e intuitivo. O *layout* é agradável, aparentemente não representando um empecilho ao usuário final.

Há farta documentação disponível, tanto oficial quanto nos fóruns, listas de discussão e artigos, apesar de a maioria ser em inglês. O próprio *site* disponibiliza documentação que abrange diversas etapas do processo de implementação do repositório, como por exemplo visão geral, instalação, administração, personalização e divulgação, entre outros.

Além disso há vários projetos de terceiros em andamento apresentados no *site* oficial. A página que trata desses projetos foi atualizada ainda neste ano e alguns dos projetos bem sucedidos apresentados foram incorporados ao DSpace, o que indica que há uma comunidade atuante e que o desenvolvimento é constante.

O *software* servidor também pode ser instalado em ambiente Windows, o que apesar de não ser encorajado por este autor nem ser o

foco principal da equipe de desenvolvimento do DSpace, pode vir a pesar na escolha de instituições com ambientes de informática heterogêneos.

Apenas uma ressalva: a documentação oficial afirma que o *Tomcat 5.x* pode ser executado com o *Java SDK 1.4* ou *1.5+*, mas não especifica que é necessário um pacote de compatibilidade para a versão *1.4*, o que pode causar problemas no momento da instalação.

4.3 CDSWare

Este *software* foi desenvolvido pela CERN, a Organização Internacional Européia para Pesquisa Nuclear e está disponível no endereço http://cdsware.cern.ch/cdsware/download.html. Segundo (PEPE, 2005), o CDSWare é:

> *"um conjunto de aplicativos que proporciona um ambiente e ferramentas para construir e gerenciar um servidor de biblioteca digital autônomo...apresenta um arquitetura modular e extensível . Cada módulo é uma entidade independente que incorpora um aspecto específico do fluxo de trabalho de uma biblioteca digital "*.

O *software* é organizado de forma que seus diversos módulos interajam entre si e com as camadas de armazenamento (banco de dados) e de interface (página *web*).

Os documentos são organizados dentro de coleções que podem ser estruturadas pelo administrador do sistema como árvores reais ou virtuais para facilitar a navegação. A submissão é feita pelo usuário e, a partir de então, o documento passa por um processo de *workflow* que pode incluir

revisão e aceitação ou não. No momento da submissão, o usuário deve informar a qual coleção o item pertence.

Todos os formatos de arquivo podem ser submetidos pelos usuários autorizados, tanto por *e-mail* quanto pela interface *web*, e o modo como esses arquivos serão exibidos é personalizável, a fim de que o *software* comporte-se de formas diferentes de acordo com o tipo de coleção. É possível definir o tamanho máximo e mínimo de arquivo a ser adicionado à obra.

A busca é realizada através de um mecanismo com sintaxe semelhante à do Google e pode combinar busca nos metadados com busca por texto integral. Além da ordenação por freqüência de palavras que permite recuperar registros similares, é incluído um método de *ranking* baseado em valores específicos de metadados. Também existe um módulo de indexação e *ranking* que permite classificar os resultados da busca pelo número de citações ou mesmo de *downloads*. A interface de busca é personalizável e já foi traduzida para 13 línguas, entre elas o português. A busca pode ser simples ou avançada e os resultados podem ser agrupados por coleção.

Usuários autorizados podem submeter obras por *e-mail* ou pela interface *Web*. Os usuários também podem criar "cestas" de documentos com o resultado de suas pesquisas e compartilhar essas cestas entre grupos de usuários, além de fazer comentários sobre documentos no repositório de forma semelhante à que acontece em *sites* que vendem livros ou CDs como o *Amazon.com*, por exemplo.

Os metadados adicionados são convertidos automaticamente para o formato de armazenamento interno do *software*, que é o MARC 21. O

MARC foi adotado por ser já um padrão bem estabelecido entre as bibliotecas, por integrar-se bem às linguagens de marcação como o XML, ser flexível o suficiente para durar por bastante tempo e adaptável a praticamente qualquer tipo de estrutura de metadados (PEPE, 2005). O MARCXML, recentemente padronizado pela *Library of Congress*, é utilizado pelo CDSWare. O esquema de metadados MARCXML pode ser utilizado em sua totalidade ou em pequenos conjuntos de elemento específicos, de acordo com a necessidade da instituição. Geralmente, quanto mais homogênea for a coleção, menor a quantidade de campos de metadados a utilizar.

Os metadados podem ser adquiridos de maneira automática (OAI-PMH). Todos os metadados incluídos no servidor são transformados para o formato nativo do CDSWare antes de ficarem disponíveis. Um dos módulos encarrega-se de fazer a tradução entre o formato nativo e outros formatos como o Dublin Core, METS ou MARC 21, no momento da exportação. No contexto da OAI, o CDSWare pode atuar tanto como provedor de dados quanto de serviços e é capacitado a fazer e atender requisições através do protocolo OAI-PMH.

Por oferecer uma ampla gama de recursos, a instalação e configuração é mais complexa que o DSpace, por exemplo. No entanto, esse esforço é requerido do administrador e não do usuário final. A relação custo x benefício acaba sendo favorável, uma vez que a dificuldade inicial é compensada com a grande flexibilidade de utilização da ferramenta.

O *software* pode, portanto, ser utilizado tanto como uma *"solução genérica de gerenciamento de documentos"* quanto como um *"sistema de*

biblioteca digital ou um repositório institucional" que visa atender a acervos de médio a grande porte (PEPE, 2005).

A documentação é farta, abrangendo visões gerais e aprofundadas do sistema, apesar de haver um erro na documentação oficial quanto à instalação que informa que o WML (*Website Meta Language*) 2.0.9 apresenta problemas de compatibilidade com o Linux Red Hat 9, quando na verdade é o 2.0.8. Também há partes do sistema que só existem na documentação, ou seja, ainda não foram completamente desenvolvidas, mas constam na documentação (por exemplo, http://cdsware.cern.ch:8000/admin/websession/guide.html, que trata do módulo de administração via *Web*), ou ainda partes do sistema para as quais a documentação não foi completada.

4.4 Nou-rau

O Nou-rau foi desenvolvido pelo Instituto Vale do Futuro em parceria com o Centro de Computação da Unicamp, é distribuído sob a licença GPL e está disponível para download em http://www.rau-tu.unicamp.br/nou-rau/. O *software* tem por objetivo implementar um sistema *online* para arquivamento e indexação de documentos, provendo acesso controlado e mecanismos eficientes para busca (NOU-RAU, 2006).

Ainda segundo (NOU-RAU, 2006), as metas do *software* Nou-Rau são:

- Armazenar qualquer tipo de documento;
- Manter informações básicas (sempre) e específicas (quando necessário) sobre cada documento;

- Permitir pesquisa tanto nos metadados quanto em texto integral;
- Permitir que se adicione ao acervo apenas o que for autorizado;
- Possibilitar a verificação de vírus externa.

A organização funcional do Nou-rau é a seguinte:

- Documentos: são arquivos submetidos ao sistema, além de seus metadados. Os documentos são armazenados e, em alguns casos, comprimidos internamente ao sistema, sendo retornados em sua forma original no momento da consulta do usuário;
- Tópicos: são agrupamentos documentos relacionados por assunto específico. Para cada tópico há um responsável por seu gerenciamento. Os tópicos podem ser organizados hierarquicamente;
- Categorias: são tipos de documentos válidos. Assim, pode-se definir que para determinado tópico, apenas documentos de determinadas categorias sejam aceitos. O tamanho máximo do documento também pode ser limitado, conforme sua categoria;
- Formatos: Cada categoria aceita um ou mais formatos de arquivo. Podem-se definir formatos além dos mais comuns, já predefinidos pelo sistema (DOC, PDF, MP3 *etc*.);

também podem ser criados grupos de formatos (por exemplo, “Todos os tipos de audio”).

Além do *software* básico necessário, para indexar o conteúdo dos documentos, é preciso instalar conversores adicionais para cada formato (DOC, PDF, TEX *etc*.).

O mecanismo de busca é implementado através da ferramenta ht://Dig. É ela que faz a indexação dos documentos a partir de informações mandadas pelo Nou-rau. Atualmente os seguintes formatos de arquivos são suportados: ASCII , HTML, RTF, SGML, WML, XML, documentos do Word, Excell e Powerpoint, PDF, Postscript, TeX, LaTeX e DVI.

No ambiente do Nou-rau, o usuário pode assumir um (ou vários) dos seguintes papéis:

- Visitante: acessa o repositório apenas para consultar;
- Colaborador: pode submeter documentos ao repositório, tornando-se o dono desses documentos;
- Responsável: administra os tópicos, inclusive aprovando ou não os documentos submetidos;
- Administrador: está hierarquicamente acima dos outros usuários, cuidando da manutenção do sistema, criação de tópicos, manutenção das categorias e formatos de documentos, bem como associando usuários às suas respectivas tarefas.

O processo de submissão de um documento envolve o *upload* por parte de um colaborador, a avaliação do responsável pelo tópico e a possível verificação de vírus pelo administrador. Se um documento não é aprovado em qualquer dessas etapas, ele é apagado da base de dados, sendo enviado um aviso ao colaborador que submeteu o documento.

A busca pode ser realizada tanto nos metadados quanto no texto integral das obras (desde que tenham sido indexadas), mas não é possível definir que se quer buscar apenas em um elemento dos metadados como o título, por exemplo. A interface do Nou-rau assemelha-se um pouco à do DSpace, sendo também de fácil utilização e bem organizada e tem uma vantagem: não é preciso trabalhar em sua tradução para começar a utilizar.

O sistema Nou-rau não oferece muita documentação. No momento em que este trabalho foi realizado, havia no *site* oficial apenas duas páginas, uma apresentando uma visão geral do sistema e outra onde eram disponibilizados o *software* em si e programas adicionais para *download*, bem como *links* para três listas de discussão – anúncios, desenvolvimento e usuários em geral. Além disso, na página de anúncios de novidades, a última notícia foi dada em 21/01/2004, liberando a versão beta3 que ainda é a disponível para *download* no *site*, o que pode indicar que, apesar de ser uma ferramenta interessante, não há interesse em levar adiante o desenvolvimento do sistema ou que, mesmo se houver, é um processo mais lento que o de outros aplicativos analisados, nos quais há grandes equipes envolvidas.

4.5 Greenstone

Segundo (GREENSTONE, 2006), o Greenstone é um *software* que visa construir e distribuir coleções de bibliotecas digitais. Foi desenvolvido e distribuído pelo Projeto Biblioteca Digital Nova Zelândia, na universidade de Waikato, em parceria com a UNESCO e a Human Info NGO, da Bélgica, e disponibilizado sob a licença GPL no endereço http://greenstone.sourceforge.net.

O ambiente de utilização divide-se em duas interfaces: uma para o usuário, disponível através do *browser*, e outra para o bibliotecário. A disponível para o bibliotecário é uma interface gráfica baseada em Java, cujo objetivo é coletar itens para o acervo, adicionar metadados, projetar as funcionalidades de navegação e pesquisa que a coleção vai oferecer para o usuário final e finalmente construir e disponibilizar a coleção. A construção de coleções também é possível a partir da interface de administração, que é uma interface *web* com menos funcionalidades que a interface Java.

Dentre os padrões de metadados comentados neste trabalho, o Dublin Core é nativamente suportado. No entanto, é possível adicionar novos padrões através de *plugins*, como por exemplo MARC e METS. Esses *plugins* também podem ser usados para submeter documentos. Para documentos textuais, há *plugins* que oferecem suporte a arquivos dos tipos PDF, PostScript, Word, RTF, HTML, texto puro, LaTeX, arquivos ZIP, Excel, Power Point, vários formatos de *e-mail* e código fonte. Para documentos multimídia há *plugins* para diversos formatos de imagem (incluindo os mais populares), MP3, Ogg Vorbis e um *plugin* genérico que pode ser configurado para outros formatos de áudio e vídeo.

Há extensa documentação em inglês, incluindo exercícios tutoriais com exemplos completos sobre como construir uma coleção a partir do zero.

A submissão é feita através da interface do bibliotecário que inclui seções para coletar os documentos (a partir da *web* ou localmente), definir os metadados para cada documento adicionado, selecionar opções da coleção (que documentos vai suportar, quais serão os índices *etc*.) e finalmente gerar a coleção e torná-la disponível na interface *web* do usuário.

Duas características que distinguem o Greenstone dos outros aplicativos analisados são a possibilidade de funcionamento em vários sistemas operacionais (inclusive não *UNIX-like*, como Windows e MAC-OSX) e a exportação de coleções para navegação local em CD-ROM, por exemplo.

4.6 Escolha do software

O esquema de metadados escolhido para representar os objetos digitais na instituição em questão foi o Dublin Core. Os principais motivos para a escolha foram:

- O nível de detalhamento dos metadados da instituição não precisa ser muito aprofundado;
- Utilização simples, o que facilita o treinamento de profissionais das mais diversas áreas;
- É o padrão básico para troca de informações via OAI (NSDL, 2005).

A utilização do padrão METS de metadados foi levada em conta. No entanto, apesar de ser ótimo para o profissional da informação pode ser complexo demais para outros usuários. Mesmo os bibliotecários precisariam de um treinamento maior que o necessário à utilização do Dublin Core, principalmente no que diz respeito aos mapas estruturais do METS.

O Greenstone e o Nou-rau foram eliminados por não atender a todos os requisitos básicos. Além disso, o Greenstone tem a particularidade de a interface do bibliotecário ser um programa em Java, não executado diretamente via *browser*, além de não proporcionar qualquer interação com o usuário. Mesmo a submissão inicial do documento teria que ser feita pelo bibliotecário. Na instituição em questão, isso geraria uma demanda impossível de atender.

Já o Nou-rau apresenta pouca documentação técnica. Por exemplo, não há nada comentado sobre a sua compatibilidade com a OAI. No entanto, num boletim publicado pelo Sistema de Bibliotecas da Unicamp em 11 de julho de 2005 o *software* já passou a ser compatível (SBU, 2005). Entretanto, a versão mais recente disponível para *download* data de 21 de janeiro de 2004. Além disso, os *links* para listas de discussão sobre o Nou-rau não estão disponíveis.

Outro ponto contra o No Nou-rau é que cada obra pode conter apenas um arquivo associado. Não seria possível, portanto, armazenar uma miniatura de uma imagem para ser visualizada no *browser* junto com a imagem mestre para ser baixada e trabalhada no computador do usuário, por exemplo. Por outro lado, uma vantagem do Nou-rau é a grande quantidade de obras em português já disponíveis nas diversas instâncias já

instaladas, uma vez que várias instituições de ensino superior brasileiras o adotaram como plataforma para a biblioteca digital.

Restaram, portanto, o DSpace e o CDSWare, ambos atendendo a todos os requisitos indispensáveis da instituição. Os grandes destaques do CDSWare são o seu poderoso mecanismo de busca, muito semelhante ao Google, e a diversidade de opções de interação com o usuário. Já o DSpace apresenta como vantagens um desenvolvimento constante, a utilização nativa de identificadores persistentes baseados no sistema *Handle,* um número muito maior de empresas e universidades que já o utilizam em ambiente de produção e várias fontes de documentação como artigos, projetos e listas de discussão. Como desvantagem, um sistema de busca relativamente simples em relação ao CDSWare, mas que não é simples em si mesmo.

Considerando o usuário alvo, no entanto, o poder do mecanismo do CDSWare atualmente seria subutilizado. Além disso, o DSpace está em constante desenvolvimento. A busca por texto integral, por exemplo, não existia em versões anteriores recentes, mas já foi implementada. Atualmente há trabalhos no sentido de exportar registros em outros formatos de metadados que não o Dublin Core. Além disso, repositórios Greenstone exportam nativamente coleções para o DSpace, independente da implementação do protocolo OAI-PMH. Considerando também que, a partir da tabela de requisitos da instituição, o DSpace apresenta mais recursos, esse *software* foi escolhido como a melhor opção.

5. A IMPLANTAÇÃO DO REPOSITÓRIO

A partir do momento da escolha do *software* foi iniciado um estudo sobre a sua aplicação. Como essa fase de estudos ainda continua, várias mudanças devem ocorrer, mas a situação atual é descrita nas seções a seguir.

5.1 Política de construção e organização do acervo

As seguintes coleções vão ser criadas inicialmente, com a hierarquia apresentada:

Materiais das disciplinas

Disciplina (uma coleção para cada disciplina)

Trabalhos de conclusão de curso

Internos

Curso (uma coleção para cada curso)

Externos

Eventos científicos

Memória histórica (composta principalmente por imagens)

Revistas & Artigos

Acta científica (revista da instituição)

Artigos em geral (artigos publicados ou não)

Academia de arte

A estrutura do DSpace adequa-se perfeitamente à configuração das coleções da forma apresentada, já que permite a hierarquia entre coleções

em múltiplos níveis. A princípio, apenas materiais já em formato digital serão incorporados à base. No entanto, há obras raras que podem ser consideradas para digitalização posteriormente. Além disso, se algum autor desejar realizar a digitalização por conta própria, nada o impede de fazer isso. Apenas a biblioteca não deverá participar do processo.

Só os arquivos de conteúdo principalmente textual, imagens, programas e combinações destes serão armazenados (aqui estão incluídos apresentações, planilhas, animações *etc*.). Entre os programas, apenas os que acompanharem alguma obra como um trabalho de conclusão de curso de computação, por exemplo, ou que tenha sido desenvolvido por algum professor para auxiliar em sua disciplina. Programas que podem ser baixados pela internet não precisam ser armazenados novamente.

É provável que mais à frente sejam criadas coleções que atendam aos setores da instituição; considerando o setor de *marketing*, por exemplo, serão então armazenados também os arquivos-fonte de cartazes, logotipos, papéis timbrados *etc*. A Academia de Arte (ACARTE), que já está sendo incluída, deve armazenar letras de músicas, arquivos em formato MID, arquivos do *software* Finale (para criação musical) e imagens com partituras digitalizadas, ficando a cargo da própria ACARTE a digitalização.

Não serão armazenados arquivos de vídeo por causa do grande consumo de espaço em disco que eles geram, bem como consumo de banda no momento do download. Além disso, já existe um departamento responsável pelo armazenamento e distribuição dos materiais em vídeo produzidos ou adquiridos pela instituição, inclusive os digitais.

Para todo formato de documento, deve ser indicado na descrição da obra o tipo de *software* que pode ser utilizado para sua leitura. Assim, o usuário final não tem que descobrir sozinho que programa poderá utilizar para abrir um documento. Trabalhos acadêmicos que os professores acharem interessante armazenar podem ser incluídos na coleção da disciplina de aplicação do trabalho.

No DSpace, um item pode aparecer em mais de uma coleção, mas sempre vai pertencer a apenas uma. No caso de obras multidisciplinares, o bibliotecário e o responsável por autorizar a submissão do trabalho ao repositório definirão a qual coleção a obra deve pertencer e em quais outras deve aparecer.

Uma observação importante é que apenas materiais sem restrições quanto a direitos autorais deverão ser armazenados. Uma vez que o objetivo é disseminar informação de forma gratuita, não faria sentido cobrar pelo acesso. Sendo assim, não é possível no momento que a instituição arque com os custos de obras com direitos autorais.

A política de construção e organização de um acervo está em constante mudança, tanto numa biblioteca convencional quanto numa digital. Portanto, como já foi citado, a descrição apresentada aqui serve apenas como base inicial para a implantação do repositório, devendo ser alterada durante o funcionamento do mesmo. É certeza que ela deve sofrer mudanças e é provável que nunca haja uma base definitiva.

Uma parte importante da organização de um acervo é definir padrão para nomes de autores e títulos. O Dublin Core define a sintaxe dos metadados, ficando a critério do catalogador inserir os dados em si. Por exemplo, o autor José de Alencar pode ser cadastrado no DSpace

como *José de Alencar* ou *Alencar, José*. Outro exemplo: duas edições de uma mesma obra devem ser objetos no repositório ou um só objeto com dois arquivos associados? Foi definido que para preencher alguns dos elementos do Dublin Core, como o *CREATOR* ou o *TITLE*, serão utilizadas as regras do padrão AACR2, já amplamente utilizadas na biblioteca convencional da instituição.

O DSpace oferece também a possibilidade de preencher alguns elementos do Dublin Core através de listas de opções pré-definidas por arquivos de configuração. Atualmente estuda-se a necessidade da utilização dessa possibilidade quanto aos assuntos das obras. A dúvida é se esse controle realmente ajuda na recuperação da informação ou se apenas desacelera o processo de submissão. Numa biblioteca convencional, listas de assuntos padronizadas eram importantes porque durante a consulta o usuário não estava diante da obra para ver seu conteúdo. No ambiente DSpace, a busca por texto integral é possível. Mesmo assim, o uso de palavras-chave é fortemente encorajado. No entanto, vocabulários controlados podem não ser absolutamente necessários no contexto da instituição. Por outro lado, limitando-se os termos permitidos, aumenta-se a precisão. Esta etapa ainda está sob estudos. Como foi citado anteriormente, o uso de vocabulários controlados também pode ser implementado em nível operacional, independendo do *software* portanto.

5.2 Políticas de acesso

É do interesse da instituição implementar o acesso aberto ao público em geral, de modo a incentivar a disseminação da produção

acadêmica. No DSpace, qualquer usuário pode fazer seu cadastro e subscrever-se para receber atualizações sobre as suas coleções de interesse. Já autores interessados em submeter obras ao repositório deverão entrar em contato com a instituição tanto para adequar-se aos procedimentos de submissão quanto para garantir que as obras submetidas ao repositório não acarretem nenhum ônus a quem quer que as utilize a partir do repositório, nem à própria instituição.

Quanto aos trabalhos de conclusão de curso especificamente, os alunos serão autorizados a submeter os seus trabalhos e os orientadores ficarão responsáveis por fazer a revisão e permitirão ou não a inclusão na base de dados. Um terceiro passo será a conferência dos metadados pelo bibliotecário, que não pode alterar o conteúdo da obra nem rejeitar sua submissão, mas pode alterar os metadados. O mesmo serve para a coleção Materiais de Disciplinas, em que haverá um professor ou mais professores responsáveis por administrar a coleção da disciplina, exceto porque a interferência do bibliotecário não será necessária neste caso. O DSpace facilita todo esse processo por implementar características básicas de *workflow*, criar usuários com acessos diferenciados e permitir utilização de perfis para grupos de usuários.

5.3 Definição dos padrões de formato e nomenclatura

Considerando o tipo de acervo pretendido na fase inicial, os arquivos acrescentados ao acervo devem ser basicamente de dois tipos:

- PDF, para arquivos de texto, apresentações, planilhas eletrônicas *etc*. Isso para garantir que o documento tenha a mesma aparência do original de onde foi gerado. O autor tem a opção de acrescentar

também os arquivos fonte, mas é preciso que se adicione pelo menos um arquivo PDF com o trabalho em sua forma integral;

- Arquivos de imagem digitalizados devem ser armazenados preferencialmente em formato JPEG, sendo opcional o armazenamento também da imagem em formato TIFF, que não oferece compressão, preservando maiores detalhes da imagem. Através dos arquivos de configuração, o DSpace permite que se opte por visualizar no *browser* uma miniatura da imagem através dos *mediafilters*, que são aplicações que trabalham a partir do conteúdo de determinados documentos – indexando texto e gerando miniaturas de imagens, por exemplo.

Quanto à nomenclatura dos arquivos submetidos, é deixado ao autor escolher a que lhe agradar mais, uma vez que o usuário pode modificar esse nome no momento de fazer o *download.* Além disso, como os arquivos são armazenados internamente na base do DSpace, não existe a preocupação com duplicidade de nomes. Mesmo na fase de exportação, cada item exportado pelo DSpace é colocado, com seus arquivos e metadados, numa estrutura de um diretório para cada item.

5.4 Questões sobre preservação digital

O objetivo desta seção é principalmente levantar questionamentos sobre a questão da preservação de objetos digitais e citar alguns dos desafios que devem ser considerados quando se pensa em repositórios digitais de informação com armazenamento a longo prazo. Segundo

(WATERS, 1996), e concordando com (LC21, 2004), são os seguintes os principais problemas da preservação de objetos digitais a longo prazo:

- Obsolescência tecnológica: as mídias para armazenamento digital podem ser frágeis e ter vida útil limitada mesmo sob condições ideais. Replicar conteúdos em outras mídias é possível, mas gera custos. Além disso, o material produzido pode ser dependente de *hardware* e *software* que não estarão disponíveis indefinidamente;
- Migração da informação digital: manter a informação digital atualizada em termos de mídia e *software,* a migração, gasta tempo e tende a apresentar erros, além de ser um processo mais complexo do que manter versões antigas de *hardware* e *software*;
- Questões legais e organizacionais: há vários tipos de licenças de direitos autorais para os diferentes tipos de obras produzidas. Também é difícil garantir que uma obra que pertença a uma coleção restrita atualmente continue restrita caso ocorra uma atualização/mudança de *software*. Também há riscos de a organização perder informações que se comprometeu a guardar ou não conseguir manter a sua integridade;
- Necessidade de grande infraestrutura: a infraestrutura citada aqui não se trata apenas da tecnológica, mas sim de uma série de conceitos que ainda estão indefinidos ou pobremente definidos. Como o mundo digital ainda é novo se comparado ao registro de informação em papel, por

exemplo, existem muitas incertezas com as quais ainda não é possível lidar facilmente.

Quanto à preservação digital, no DSpace há três *status* que um arquivo pode assumir (DSPACE SYSTEM DOCUMENTATION, 2005):

- Suportado: o tipo de arquivo é conhecido e a instituição garante que será possível visualizar seu conteúdo a longo prazo;
- Conhecido: a instituição reconhece e armazena o tipo de arquivo, pretende obter informações suficientes para transformá-lo para o status de suportado;
- Desconhecido: a instituição não reconhece o tipo de arquivo, mas mesmo assim compromete-se a preservá-lo.

Claro, isso não representa solução, apenas uma ferramenta básica para tentar posicionar a instituição quanto aos formatos de arquivos presentes na base de dados. Cada biblioteca busca suas próprias soluções e, até o momento, há apenas estratégias mais ou menos adequadas a cada caso. Como conclusão, (WATERS, 1966) apresenta o seguinte:

- A principal frente de defesa contra a perda de material digital é diretamente com os autores ou organizações onde a obra foi produzida;
- Um ponto crítico da infraestrutura de preservação é a existência de múltiplas instituições de confiança capazes de armazenar, migrar e disponibilizar material digital;

- É preciso criar mecanismos de certificação para os repositórios digitais para gerar um ambiente confiável.

(PADI, 2006) oferece algumas estratégias para a preservação digital a longo prazo, além das já citadas:

- Padrões: a instituição deve utilizar padrões abertos e já bem estabelecidos em vez de formatos proprietários;
- Emulação: reprodução de ambientes de *software* e *hardware* que traduzam código de um sistema de computação para outro;
- Encapsulamento: como parte da estratégia de emulação, objetos e seus metadados armazenados juntos para ajudar na renderização do objeto posteriormente;
- Metadados de preservação: metadados que descrevam os requisitos para a visualização do objeto digital posteriormente.

Mais uma vez, a idéia aqui é apenas levantar questões para reflexão. Uma análise mais profunda da instituição, do conteúdo que ela deseja armazenar, do tipo de usuário que possui, enfim, de todo o contexto em que o repositório digital estará inserido é necessária antes de estabelecer quaisquer procedimentos e deve-se manter em mente que, devido à relativa imaturidade da produção de obras puramente em meio digital, esses procedimentos certamente deverão ser reavaliados com freqüência.

5.5 Dificuldades encontradas

A instalação e configuração dos aplicativos que servem de plataforma para um repositório digital de informações é relativamente simples, não apresentando maiores problemas para o administrador de rede. No DSpace o processo de criação das coleções é tão ou mais simples que os outros aplicativos analisados. De todos, o Greenstone é que apresenta a interface mais complexa para criação de coleções. Durante a realização deste trabalho, os principais problemas técnicos decorreram de erros na documentação dos aplicativos mais do que na instalação em si.

A grande dificuldade é realmente elaborar o contexto em que o repositório de informações vai ser inserido, bem como definir o nível de descrição desejado para os objetos digitais. Questões como que tipo de material deverá ser armazenado e por quanto tempo, qual o formato e nomenclatura dos arquivos, quem terá acesso e a que partes do sistema, como agrupar as pessoas e obras na estrutura do repositório, entre outras, formam as partes mais difíceis do processo dessa implantação.

Apesar da dificuldade, só a partir desse levantamento preliminar, ainda que normalmente muito sujeito a mudanças, é possível realizar a análise direcionada dos aplicativos disponíveis e finalmente chegar a uma escolha consciente de um *software* que possa atender às necessidades da instituição no contexto da implantação do repositório.

6. Conclusão e projetos futuros

A partir do levantamento dos requisitos da instituição e da análise dos principais aplicativos atualmente disponíveis, este autor conclui que existe *software* livre em quantidade e qualidade suficientes para atender às necessidades de diversos tipos de instituições quanto aos repositórios digitais de informação. Segundo (BARTON, 2005), são vários os motivos que levam as pessoas a utilizar os repositórios institucionais. Entre elas:

- Comunicação entre universidades;
- Armazenamento de materiais de aprendizado;
- Publicação eletrônica;
- Gerenciar coleções de documentos de pesquisa;
- Preservar materiais a longo prazo;
- Hospedar conteúdo digitalizado;

A análise de alguns dos aplicativos disponíveis demonstrou que para qualquer dessas necessidades, já existem ferramentas em plenas condições de utilização. Além disso, a filosofia do *software* livre e a organização modular do sistema operacional GNU/LINUX permitem que novas funcionalidades sejam implementadas, seja através do desenvolvimento colaborativo das comunidades de usuários de cada *software*, seja pela combinação de várias ferramentas existentes a fim de atingir determinado objetivo.

Quanto ao futuro, ainda durante a implantação efetiva do repositório digital serão estudadas as estratégias mais eficientes de

backup dos dados. Como qualquer sistema que envolva informações importantes, em caso de falha de *hardware*, *software* ou mesmo no caso de um *upgrade* na plataforma do repositório, é necessário ter cópias de segurança do conteúdo da base de dados.

Além disso, atualmente existe um servidor utilizado para ensino à distância que funciona em conjunto com um sistema de alta disponibilidade. Será estudada a possibilidade de fazer o mesmo com o servidor do repositório, de forma a garantir que este esteja em funcionamento e protegido contra falhas durante o maior tempo possível.

REFERÊNCIAS BIBLIOGRÁFICAS

AD HOC WORKING GROUP. *Dublin Core Element*: COVERAGE. Disponível em <http://www.alexandria.ucsb.edu/public-documents/metadata/dc_coverage.html>. Acesso em: 06 abr. 2006

ALVARENGA, Lídia. *Representação do conhecimento na perspectiva da Ciência da informação em tempo e espaços digitais.* Revista eletrônica de biblioteconomia e ciência da informação. n. 15, 2003. Disponível em: <http://www.encontros-bibli.ufsc.br/Edicao_15/alvarenga_representacao.pdf >. Acesso em: 13 mar. 2006.

BARTON, Mary R. *Creating an institutional repository*: LEADIRS workbook. Cambridge : MIT Libraries, 2005. Disponível em <http://www.DSpace.org/implement/leadirs.pdf>. Acesso em: 12 abr. 2006

BERGMAN, Michael K. *The deep web:* surfacing hidden value. The Journal of Electronic Publishing, v.7, n.1, Ago. 2001. Disponível em <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em: 14 abr. 2006.

BRANTON, Ann; CHEN-GAFFEY, Aiping. *MARC 21 Tutorial.* University of Southern Mississippi; Slippery Rock University, 2004. Disponível em

<http://www.lib.usm.edu/~techserv/pdc/marc21_tutorial_ie/marcintroIE.htm>. Acesso em: 06 abr. 2006

COLE, Timothy W. *Creating a Framework of Guidance for Building Good Digital Collections.* First Monday Journal, v.7, n..5. maio de 2002 Disponível em <http://firstmonday.org/issues/issue7_5/cole/index.html>. Acesso em: 12 abr. 2006.

DCMI. *Dublin Core Metadata Element Set version 1.1:* reference description. Dublin Core Metadata Initiative, 2004. Disponível em <http://dublincore.org/documents/dces/>. Acesso em 06 abr. 2006.

DSPACE INSTANCES. Cambridge : Massachusetts Institute of Technology; Palo Alto : Hewlett-Packard Company, 2005. Disponível em: <http://wiki.DSpace.org/DSpaceInstances> . Acesso em: 14 abr. 2006.

DSPACE SYSTEM DOCUMENTATION. Cambridge : Massachusetts Institute of Technology; Palo Alto : Hewlett-Packard Company, 2005. Disponível em: <http://DSpace.org/technology/system-docs> . Acesso em: 14 abr. 2006.

EDUCADVENTISTA.ORG. *Nossa história.* Disponível em: <http://www.educadventista.org.br/index.php?option=com_content&task=view&id=13&Itemid=39>. Acesso em: 15 abr. 2006.

FURLAN, José Davi; IVO, Ivonildo da Motta. *Megatendências da tecnologia da informação.* São Paulo : Makron Books, 1993.

GARTNER, Richard. *METS:* Metadata Encoding and Transmission Standard. Oxford University Library Services, 2002. Disponível em <http://www.jisc.ac.uk/uploaded_documents/tsw_02-05.pdf>. Acesso em: 08 abr. 2006.

GORMAN, Michael (coord.); WINKLER, Paul W. (coord). *Código de Catalogação Anglo-Americano Segunda Edição.* The American Library Association, 1983.

GREENSTONE. *The Greenstone digital library software*. Disponível em <http://www.greenstone.org>. Acesso em 28 mar. 2006.

HANSEN, Preben. *User Guidelines for Dublin Core Creation.* Nordic Metadata Project, 1999. Disponível em: <http://www.sics.se/~preben/DC/DC_guide.html>. Acesso em: 07 abr. 2006

KNIGHT, Gareth. *An introduction to metadata requirements for an e-print repository*. SHERPA Arts & Humanities Data Service, 2004. Disponível em <http://www.sherpa.ac.uk/documents/D2-6_Report_on_Metadata_Issues.pdf >. Acesso em: 14 abr. 2006.

LC21: A digital strategy for the Library of Congress. Washington : National Academy Press, 2004. Disponível em <http://books.nap.edu/catalog/9940.html>. Acesso em: 06 mar. 2006.

LIBRARY OF CONGRESS. *MARC Standards.* The Library of Congress – Network Development and MARC Standards Office, 2005. Disponível em <http://www.loc.gov/marc/>. Acesso em: 15 abr. 2006.

LIBRARY OF CONGRESS. *METS*: Introdução e tutorial. Washington : The Library of Congress, 2005. Disponível em <http://www.loc.gov/standards/mets/METSOverview.v2_port.html>. Acesso em: 10 mar. 2006.

LIBRARY OF CONGRESS HELP DESK. *Understanding MARC Bibliographic.* Washington : The Library of Congress, 2003. Disponível em <http://www.loc.gov/marc/umb/um01to06.html>. Acesso em: 09 mar. 2006.

LYNCH, Clifford. *Institutional repositories:* essential infrastructure for scholarship in the digital age. ARL Bimonthly Report, n. 226, fev. 2003. Disponível em: <http://www.arl.org/newsltr/226/ir.html>. Acesso em: 11 mar. 2006.

NELSON, Michael L. et al. *mod_oai:* An Apache module for metadata harvesting. Norfolk : Old Dominion University; Los Alamos : Los Alamos National Laboratory, 2004 Disponível em <http://arxiv.org/ftp/cs/papers/0503/0503069.pdf>. Acesso em: 12 abr. 2006.

NISO (National Information Standards Organization). *Understading Metadata* Bethesda : NISO Press, 2004. Disponível em

<http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>. Acesso em: 08 mar. 2006

NOERR, Peter. *The Digital Library Toolkit.* 3.ed. Sun Microsystems, 2003. Disponível em <http://www.sun.com/products-n-solutions/edu/whitepapers/digitaltoolkit.html >. Acesso em: 06 mar. 2006

NOU-RAU. *Descrição do Nou-Rau*. Unicamp; Instituto Vale do Futuro. Disponível em <http://www.rau-tu.unicamp.br/nou-rau/desc-pt.html> Acesso em: 11 mar. 2006.

NSDL. *OAI Best Practices*. The National Science Digital Library, 2005. Disponível em <http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?MultipleMetadataFormats>. Acesso em: 14 abr. 2006.

OAI (Open Archives Initiative). *The Open Archives Initiative Protocol for Metadata Harvesting.* 2004. Disponível em <tttp://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>. Acesso em: 14 abr. 2006.

OPEN ARCHIVES FORUM. *OAI For Beginners.* University of Bath, 2003. Disponível em <http://www.oaforum.org/tutorial/english/intro.htm>. Acesso em: 29 mar. 2006.

PADI (Preserving Access to Digital Information). *Digital preservation strategies.* National Library of Australia. Disponível em http://www.nla.gov.au/padi/topics/18.html>. Acesso em : 12 abr. 2006.

PEPE, A et al. *CERN Document Server Software*: the integrated digital library. CERN : Geneva, 2005. Disponível em <http://cdsware.cern.ch/cdsware/doc/elpub2005.pdf>. Acesso em: 06 abr. 2006

PRASAD, A.R.D.; GUHA, Nabonita. *Interoperability and the OAI-PMH.* Bangalore : DRTC-HP International Workshop on Building Digital Libraries with DSpace, 2005. Disponível em <https://drtc.isibang.ac.in/handle/1849/245>. Acesso em: 08 abr. 2006.

RFC2046*:* Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. The Internet Engineering Task Force, 1996. Disponível em: <http://www.ietf.org/rfc/rfc2046.txt> . Acesso em: 15 abr. 2006.

SBU. *Biblioteca digital de teses da UNICAMP, a maior do Brasil*. Sistema de Bibliotecas da UNICAMP, 2005 Disponível em <http://143.106.108.14/BoletimSBU/2005/julho/noticias/libdigi.php>. Acesso em: 15 abr. 2006

UNASP. *O UNASP – Histórico.* Centro Universitário Adventista de São Paulo, 2006. Disponível em: <http://www.unasp.edu.br/o-unasp-historico.html>. Acesso em 15 abr. 2006.

W3C. *Overview of SGML Resources.* World Wide *Web* Consortium, 2004. Disponível em <http://www.w3.org/MarkUp/SGML/>. Acesso em: 15 abr. 2006.

W3C. *Extensible Markup Language (XML)*. World Wide *Web* Consortium, 2006. Disponível em < http://www.w3.org/XML/>. Acesso em: 15 abr. 2006.

WATERS, Donald; GARRET, John.. *Preserving Digital Information:* Report of the Task Force on Archiving of Digital Information. Washington : 1996. Disponível em <http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf>. Acesso em 06 abr. 2006.

APÊNDICE A

A tabela a seguir apresenta os aplicativos analisados e os programas necessários ao seu funcionamento

Aplicativo	Versão	Sist. Op.	Serv. Web	Adicionais	Banco de dados
DSpace	1.3.2	*UNIX-like* *Windows*	Apache Tomcat	Apache Ant Java SDK	PostgreSQL ou Oracle
Greenstone	2.70	*UNIX-like* Windows MAC-OSX	Apache	- PERL - Java SDK - Plugins para padrões de metadados e formatos de arquivos adicionais	GDBM (GNU Database Manager)
Nou-rau	beta 3	*UNIX-like*	Apache	- PHP - FILE - HTDIG - PERL - Conversores de formatos de arquivos (para indexação *full-text*)	PostgreSQL

Aplicativo	Versão	Sist. Op.	Serv. Web	Adicionais	Banco de dados
CDSWare	0.7.1	*UNIX-like*	Apache	- Python - PHP - WML - *Parsers* para XMLMARC (opcional) - Gnuplot (recomendado) Implementação de Commom LISP (CLISP, SBCL ou CMUCL. (recomendado) - Conversores de formatos de arquivos (para indexação *full-text*)	MySQL