MAC5701
TÓPICOS EM COMPUTAÇÃO

# BIBLIOTECAS DIGITAIS

Evelyn Cristina Pinto
Orientador Prof. Dr. Imre Simon

1 Introdução ..... 3
Questões políticas ..... 3
2 Quem é a favor e quem é contra? ..... 3
3 *Peer review* eletrônico ..... 6
4 *Open Archives Initiative* (OAI) ..... 6
Questões técnicas ..... 7
5 O Protocolo OAI ..... 7
6 Provedores de Dados ..... 8
7 Provedores de Serviços ..... 8
8 *Harvesters* ..... 8
9 Técnicas de implementação ..... 8
9.1. SODA: *Smart Objects, Dumb Archives* ..... 8
10 Conclusão ..... 9
Bibliografia ..... 11

# BIBLIOTECAS DIGITAIS

## 1 Introdução

Devido ao crescente processo de digitalização de documentos tais como artigos científicos, jornais, revistas, teses, entre outros, e à motivação do acesso rápido e eficiente a esses documentos, surgiu a necessidade de armazená-los de forma organizada. Para tanto, as bibliotecas digitais foram desenvolvidas.

As bibliotecas digitais são repositórios de documentos digitalizados, tais como arquivos de formato PDF, TXT, GIF, JPG, DOC, etc, ou repositórios de informações descritivas de documentos (digitalizados ou não) como título do documento, assunto, autores, número de páginas, formato, onde se encontra disponível, etc. Este tipo de informação é comumente chamado de metadado.

Entre as bibliotecas virtuais famosas na Internet, destacam-se arXiv do Laboratório de Los Alamos, que promoveu esse sistema numa iniciativa inovadora de abrir para o público artigos científicos excelentes antes mantidos escondidos. Também merece ser citado o repositório virtual da BiomedCentral (revista científica da área biomédica), além da biblioteca digital de teses do MIT.

# Questões políticas

## 2 Quem é a favor e quem é contra?

A tecnologia digital está tornando mais fácil escrever livros, salvar seu conteúdo, e, de fato, mais fácil salvar qualquer coisa escrita. A reprodução eletrônica é quase instantânea e muito barata. A distribuição de documentos digitais é bem fácil e pode atingir um público enorme e desconhecido. Isto implica que mais e mais informação estará disponível e que a habilidade para se encontrar informação é crescentemente importante; a possessão dela será menos importante. As bibliotecas vão achar mais barato e mais fácil armazenar eletronicamente a informação.

A tecnologia pode mover-nos para poucos recursos ou para muitos, dependendo de como nós a aplicamos. O mesmo argumento que Benjamin Franklin usou pra justificar o serviço postal nós podemos aplicar às bibliotecas digitais: para o bem da democracia, os cidadãos precisam de acesso à informação [11].

## 2.1. Oráculo Bibliográfico: uma biblioteca digital completa

Alguns pesquisadores vêm sonhando com uma biblioteca de literatura científica completa facilmente disponível [12]. A Internet tem se mostrado capaz de fornecer infra-estrutura para que esse sonho se realize. Com a rápida evolução e aperfeiçoamento dos métodos sintáticos de indexação e os de ordenação por relevância, os motores de busca na Rede são capazes de responder quase instantaneamente a muitas perguntas.

Recentemente, surgiu um serviço muito útil aos pesquisadores da área da Computação, o ResearchIndex ou CiteSeer. Ele procura constantemente na Internet artigos científicos na área de Ciência da Computação. Ao encontrar, o CiteSeer indexa todas as palavras que aparecem nele. Utiliza as referências bibliográficas citadas no artigo para estabelecer um índice de relevância para os artigos catalogados. Assim, permite medir o impacto de um dado artigo na literatura científica da Ciência da Computação. Este serviço já mostra a viabilidade de uma biblioteca digital completa, ou seja, uma biblioteca na Internet que abranja todos os ramos da ciência.

## 2.2. Problemas

Mas em 1964, Arthur Samuel predisse que em 1984, as bibliotecas "em papel" iriam desaparecer, exceto nos museus. Por que isso não aconteceu? Uma das razões é a de nós não encontramos facilmente 3 bilhões de dólares para converter 100 milhões de livros à forma eletrônica, mais o adicional de soma muito maior para compensar os direitos autorais da maior parte desses livros. As economias das bibliotecas digitais estão diretamente relacionadas a da publicação digital, das revistas científicas e da indústria de rede, de questões de segurança e fraude [11]. É relativamente fácil adulterar um texto ou depositar um documento numa biblioteca virtual em nome de outra pessoa.

Também há o problema de garantir acesso àqueles que não possuem computadores ou serviços *online* e a nossa sensível relutância em mudar um sistema que funciona para outro desconhecido. Para alguns bibliotecários, digitalizar livros é um exemplo de consertar algo que não está quebrado. Além disso, alguns leitores de livros estão acostumados a folhear papéis para fazer sua leitura e ainda preferem fortemente esse método a ler um livro numa tela de computador [11].

Peter Lyman, um ilustre pesquisador da Biblioteconomia, argumenta que as bibliotecas virtuais, por estarem numa rede de computadores esparramados pelo mundo todo, tiram os livros das prateleiras das bibliotecas das universidades (onde existe uma cultura e contexto para se entender o material), e transportam-nos

para espaços geográficos e culturais completamente diferentes [1]. Dessa forma, o entendimento do conteúdo pode ser complicado.

Ainda como força contrária à difusão da informação aberta, a China e Singapura são exemplos de países que tentam controlar o espalhamento da informação eletrônica. Mas a mídia eletrônica é difícil de controlar, e é mais privada do que governamental.

## 2. 3. Propriedade Intelectual

É difícil medir o quanto cada um dos aspectos citados acima têm influenciado a evolução das bibliotecas digitais. Porém, pode-se perceber que a instituição da propriedade intelectual tem uma *grande* força contrária a esse progresso. Para uma rápida demonstração da força da propriedade intelectual, gostaria de destacar as recentes disputas entre SCO/Caldera (portadora dos direitos de propriedade intelectual de um software da família UNIX) e a IBM.

A SCO/Caldera acusa alguns desenvolvedores do Linux, patrocinados pela IBM, de haverem copiado parte do código fonte do *kernel* (núcleo do sistema operacional) UNIX para o *kernel* do Linux [14]. O Grupo SCO, avaliado em vinte milhões de dólares, está agora cobrando da IBM uma indenização de um bilhão de dólares.

No caso das bibliotecas digitas, temos as revistas científicas, que também faturam bilhões de dólares por ano. Elas estabeleceram uma prática onde o autor transfere os seus direitos autorais para a entidade que publica o seu trabalho. A partir deste momento, o editor passa a ter uma enorme influência sobre a disseminação do trabalho. E assim, o autor não pode depositar seu trabalho numa biblioteca digital aberta ao público. Há também outras revistas científicas que adotaram a prática de recusar a publicação de artigos que circularam muito nos repositórios de *preprint* [13].

# 3 *Peer review* eletrônico

Temores sobre os possíveis prejuízos ao sistema clássico de *peer review* estão atrapalhando a evolução da comunicação na universidade. Em particular, eles estão inibindo o desenvolvimento do acesso gratuito a artigos nas bibliotecas digitais. Porém, novas formas de comunicação oferecem maior liberdade de escolha aos pesquisadores, e provêem um mecanismo de *feedback* da informação disponível de forma mais rápida, completa e flexível.

Algumas das críticas que podemos fazer ao sistema tradicional de *peer review* são: ele não garante *corretude* nem *novidade* dos resultados, mesmo entre os mais seletos e prestigiados jornais, além de atrasar a publicação de um artigo.

## 3.1. *Peer review* "instantâneo"

Um exemplo de como o *peer review* eletrônico pode ser eficiente é o caso dos três indianos e o teste do número primo [2]. Estes pesquisadores provaram que o teste "*se um número natural é primo*" pode ser feito em tempo polinomial. Este era um velho e famoso problema em aberto da Matemática e Ciência da Computação.

No domingo, dia 4 de agosto de 2002, Maninda Agrawal, Neeraj Kayal e Nitin Saxena do Instituto Indiano de Tecnologia em Kanpur enviou um artigo com a espantosa prova a vários especialistas reconhecidos em teste de números primos. Na terça, 6 de agosto, os autores então colocaram o artigo nos seus sítios e enviaram e-mails anunciando o resultado. Na quinta, 8 de agosto, o New York Times publicou o caso.

## 3.2. Futuro do *peer review*

Nós podemos esperar um sistema muito mais heterogêneo evoluir. Por exemplo, a atual proibição de publicação simultânea em múltiplos jornais será provavelmente descartada como uma relíquia da era Gutenberg, quando os recursos de imprensa eram escassos. Sem dúvida, essa iniciativa favorecerá muito o desenvolvimento das bibliotecas digitais.

# 4 *Open Archives Initiative* (OAI)

A questão da interoperabilidade é um dos problemas mais significativos que as bibliotecas digitais atualmente enfrentam. A inabilidade para unificar, filtrar e prover serviços de conteúdo remoto úteis limita-as a cobrirem apenas as

demandas locais. A OAI é uma organização de esforço para embutir interoperabilidade técnica entre repositórios distribuídos. O objetivo da OAI é facilitar a descoberta de conteúdo em repositórios distribuídos [16].

# Questões técnicas

## 5 O Protocolo OAI

O Protocolo OAI é simplesmente uma interface em que um servidor de rede (não necessariamente um servidor de *e-print*) pode empregar para que os metadados descritivos de objetos residentes no servidor estejam disponíveis para aplicações externas que desejem coletar estes metadados.

Ele utiliza um conjunto muito simples de ferramentas de transações *request-response* baseadas no protocolo HTTP, para comunicação entre um coletor (harvester) e um repositório [5]. O servidor retorna uma série de conjuntos de elementos: metadados (em XML) mais identificadores (por exemplo, URLs) . O protocolo OAI-PMH provê um conjunto de ferramentas inter-operacionais independente de aplicação, baseado na "*colheita" de dados (metadata haversting).*

Nesse protocolo são utilizados identificadores únicos que determinam um item dentro de um repositório. Um item é uma espécie de recipiente que armazena metadados sobre um documento digital que pode estar em múltiplos formatos (PDF, DOC, TXT, JPG, etc).

O protocolo OAI é composto de seis verbos de requisição. São eles:

- Identify – traz em formato XML as principais informações do repositório de dados como nome, identificador, *e-mail* do administrador do repositório, etc.
- ListMetadataFormats – lista os formatos de metadados implementados pelo sistema do repositório; pelo protocolo OAI, é obrigatória a implementação de pelo menos o formato Dublin Core.
- ListSets – traz a árvore de assuntos que classificam os documentos no repositório.
- ListIdentifiers – lista todos identificadores de item do repositório.
- ListRecords – lista todos itens do repositório.
- GetRecord – dado um identificador de item, lista o item correspondente.

Com as suas ferramentas, o protocolo viabilizou aplicações que visam sincronização de repositórios, busca numa federação de repositórios, troca de metadados entre bibliotecas digitais, entre outros.

# 6 Provedores de Dados

Os provedores de dados são sistemas que dão suporte ao protocolo OAI-PHM como um meio de expor metadados. Por exemplo, no IME-USP, em http://eprints.ime.usp.br, há um servidor de artigos digitais implementado através do software EPrints, da Universidade Southampton. Outro exemplo é o repositório Open Vídeo (http://www.open-video.org), que contém imagens digitalizadas e expõe os metadados desses arquivos via protocolo OAI.

# 7 Provedores de Serviços

Os provedores de serviços fazem uso dos metadados coletados via protocolo OAI-PHM como base para construção de serviços. Um bom exemplo de provedor de serviço é o ARC, da Old Dominium University. Outro é o DP9 , um portal que permite a indexação de um provedor de dados OAI pelos motores de busca da Internet [10]. Isso torna as páginas do provedor de dados acessíveis aos motores de busca que não implementam o protocolo OAI.

# 8 *Harvesters*

O ARC [17] é o primeiro serviço de busca federada baseado no protocolo OAI. Ele se originou do protótipo UPS (Universal Preprint Service), que foi desenvolvido para um estudo da viabilidade da construção de um serviço de busca cruzada entre repositórios [9]. O UPS continha aproximadamente duzentos mil registros coletados de seis repositórios. Uma vez que o protocolo se estabilizou, foi possível transportar as idéias do UPS para o ARC, com uma performance mais elevada na capacidade de busca e conteúdos mantidos atualizados através do protocolo OAI.

Outro *harvester* famoso é o NDLTD (Networked Digital Library of Theses and Dissertations), um coletor metadados de bibliotecas digitais que armazenam teses e dissertações [15]. Atualmente, ele agrega cento e cinqüenta e nove bibliotecas virtuais de teses universitárias, inclusive as da UNESP, USP, PUC-RJ e PUC-MG.

# 9 Técnicas de implementação

## 9.1. SODA: *Smart Objects, Dumb Archives*

Muitas das funções tradicionalmente associadas aos provedores de dados (*archives)* foram direcionadas para dentro dos *buckets*, resultando em objetos mais "espertos" e repositórios (*archives*) mais "burros". *Buckets* são uma classe especial de objetos digitais, agentes agregativos e inteligentes. São também protocolos independentes da biblioteca digital. Eles provêem um mecanismo de agrupamento de objetos de informações relacionadas numa única entidade lógica para armazenamento. O modelo SODA surgiu da necessidade do laboratório NASA Langley Research Center de agrupar num mesmo local documentos armazenados de formas diferentes [7].

Esse laboratório produz muitos relatórios em formato PS, PDF ou DOC, mas freqüentemente há uma planilha ou um software anexado ao documento. O grupo de pesquisa tinha servidores diferentes para armazenar os softwares e os relatórios. Por isso, após alguns anos, com algumas transformações no ambiente do servidor, os índices de relacionamento dos anexos se perdiam. Para solução desse problema, surgiram os *buckets*, que são objetos que guardam juntamente todas as referências para os documentos e softwares relacionados a um mesmo relatório.

### 9.2. INTERFACES

No sítio da OAI , podemos encontrar dois *softwares* de interface que são: OAICat e OAIHarvester [6]. O OAICat é uma aplicação para a *web* baseada em Java Servlet e provê um *framework* para implementação de repositórios. O *framework* pode ser adaptado para trabalhar com tipos muito diferentes de repositórios de dados, através da implementação de algumas interfaces Java. Já o OAIHarvester é uma interface para implementação de *harvesters*.

Como podemos notar, tanto no modelo SODA quanto nas interfaces descritas acima, a OAI incentiva a aplicação de técnicas como encapsulamento, flexibilidade, reuso de interfaces e funcionalidade, e principalmente, uma separação clara de funcionalidade entre *provedores de dados* e de *provedores de serviços*.

## 10 Conclusão

Muitos pesquisadores imaginaram um crescimento exponencial e acelerado do volume de documentos depositados nas bibliotecas digitais. Porém, a evolução não está tão rápida como se esperava. Por quê? Minha dissertação de mestrado trará da seguinte forma algumas respostas a essa pergunta:

a. Utilizando os metadados do ARC, gerarei gráficos que mostrem quantos documentos foram depositados num certo espaço de tempo. Desses

documentos, quantos deles têm seu conteúdo disponível em alguma biblioteca digital. Além disso, é muito interessante a análise de quantos documentos sobre determinado assunto foram publicados. Para determinar o assunto de um documento, implementarei algoritmos de extração de tópicos;

b. Farei uma análise do impacto da propriedade intelectual na evolução das bibliotecas digitais. Para tanto, estudarei principalmente a economia e política dos direitos autorais entregues às revistas científicas.

Outra lacuna na área das bibliotecas digitais é a falta de um identificador único para documentos digitais que seja legível para humanos. Os identificadores que existem hoje são baseados na *URL-base* mais um identificador numérico interno à biblioteca digital.

# Bibliografia

1. Peter Lyman. *Scholarly Communication and Technology.* Conferência sobre Documentos Digitais e o Futuro da Comunidade Acadêmica. Emory University. 1997.
2. Andrew Odlyzko. *Peer and Non-peer Review.* http://www.dtc.umn.edu/~odlyzko/doc/peer.review.txt.
3. Ian Parberry. *A Guide for New Referees in Theorical Computer Science*. Universidade de Norte do Texas. 1994.
4. Open Archives Initiative – Guia de Repositório. http://www.openarchives.org/OAI/2.0/guide-repository.htm.
5. Protocolo OAI. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.
6. Ferramentas e Serviços para Repositórios OAI. http://www.openarchives.org/tools/tools.html.
7. Michael L. Nelson, Kurt Maly, Mohammead Zubair, Stewart N. T. Shen, *SODA: Smart Objects, Dumb Archives*. Conferência sobre Pesquisa e Tecnologia Avançada para Bibliotecas Digitais. Paris. 1999.
8. Lista de discussão OAI-general. Subject: *OAI and intellectual property issues*.
9. Xiaoming Liu, Kurt Maly, Mohammad Zubair, Michael L. Nelson. *Arc – An OAI Service Provider For Cross-Archive Searching*. Old Dominium University.
10. Xiaoming Liu, Kurt Maly, Mohammad Zubair, Michael L. Nelson. *DP9: An OAI Gateway Service For Web Crawlers*. Old Dominium University.
11. Michael Lesk. *Practical Digital Libraries, Books, Bytes & Bucks*. 1997. Editora Morgan Kaufmann. San Francisco, California. 1-8.
12. Imre Simon. Oráculo Bibliográfico: Sonhos de Um Pesquisador, 2002. http://www.ime.usp.br/~is/papir/oraculo/oraculo-rusp/oraculo-rusp.html
13. Debate na Revista Nature sobre Acesso Aberto. http://www.nature.com/nature/debates/e-access/Articles/tenopir.html.
14. Eric Raymond. OSI Position Paper on the SCO-vs.-IBM Complaint. http://catb.org/~esr/hackerlore/sco-vs-ibm.html.
15. Networked Digital Library of Theses and Dissertations. http://www.ndltd.org/.
16. Open Archives Iniative http://www.openarchives.org
17. ARC http://arc.cs.edu.odu