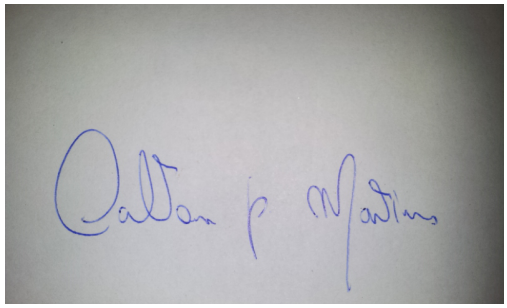
# Relatório do projeto de arquitetura em rede para integração federada

**Universidade Federal de Goiás**
Faculdade de Informação e Comunicação
Laboratório de Políticas Públicas Participativas

**Ministério da Cultura**
Coordenação Geral de Cultura Digital

**Projeto: "Laboratório de Políticas Públicas Participativas: foco em acervos digitais"**
Setembro/2015

Responsável pelo relatório



_____

Prof. Dr. Dalton Lopes Martins
Coordenador do Laboratório de Políticas Públicas Participativas

# Sumário

# 1. Introdução

O presente relatório tem por objetivo apresentar e colocar em análise as principais práticas e elementos que devem ser considerados para a integração federada de repositórios de documentos digitais em rede, entendendo que para isso é preciso colocar em discussão o modelo de comunicação a ser adotado, os padrões técnicos que devem ser integrados e as práticas de gestão da informação que precisam ser consideradas para que o plano técnico dos sistemas entregue informação de qualidade para a produção de novos serviços oriundos de seu potencial de integração. Para isso, faremos um breve histórico do movimento Open Archives Initiative[1], procurando ressaltar ao longo do relatório como esse movimento não reflete apenas uma preocupação com padrões técnicos, mas sim procura oferecer resposta a um modelo de comunicação em rede para interoperabilidade entre repositórios, que pode e deve ser considerado como etapa inicial de produção de redes de acervos digitais no âmbito da estratégia do Ministério da Cultura. O relatório, portanto, fundamenta as bases técnicas e conceituais da arquitetura de rede utilizada pelo projeto **"Laboratório de Políticas Públicas Participativas: foco em acervos digitais"** parceria entre a Universidade Federal de Goiás e o Ministério da Cultura.

De partida, vale dizer que o movimento Open Archives Initiative vem se estabelecendo como um modelo de transporte e compartilhamento de metadados[2] desde a publicação do protocolo OAI-PMH (Open Archives Initiative Protocolo for Metadata Harvesting)[3] em janeiro de 2001. Sendo um modelo de arquitetura da informação projetado para ampliar a interoperabilidade entre bibliotecas digitais e facilitar a disseminação da informação de forma mais eficiente (Cole e Foulonneau, 2007, p.3), tem sido utilizado como base no desenvolvimento de novos serviços de dados para essas bibliotecas.

A produção de novos serviços presume a possibilidade de agregação dos dados a partir de normas e convenções básicas compartilhadas entre as bibliotecas digitais que se deseja integrar. Uma vez respeitadas e implementadas essas normas e convenções básicas, é necessário analisar a qualidade semântica dos dados coletados, permitindo avaliarmos as reais possibilidades de agregação e representatividade desses dados. Procedimentos de normalização e tratamento também são elementos fundamentais a serem considerados na melhoria das condições de agregação dos dados. Vale dizer que com isso se abre todo um novo plano de funções técnicas e sociais no âmbito da cultura digital, o que chamaremos ao longo das experimentações desse projeto pela expressão "curadoria digital". Como veremos, o uso do protocolo OAI-PMH tem incentivado a produção de novos serviços e facilitado esses procedimentos de tratamento e integração da informação.

1 Iniciativa dos Arquivos Abertos
2 Metadados: informação estruturada utilizada para descrever um recurso de informação em particular
3 Protocolo para Coleta de Metadados da Iniciativa dos Arquivos Abertos

Há um crescimento expressivo no número de bibliotecas digitais que ofertam metadados de seu conteúdo seguindo os padrões do protocolo OAI-PMH (Cole e Foulonneau, 2007, p.55) que envolve diversas instituições, dentre elas universidades, centros de pesquisa, laboratórios, bibliotecas e serviços especializados na disponibilização de produções científicas ao redor do mundo. No Brasil, o movimento segue a mesma tendência, sendo que em 2009 já tínhamos 227 títulos de revistas em formato eletrônico registradas no Portal Scielo[4], 757 revistas registradas no IBICT[5] e 90 bibliotecas de teses e dissertações usando o protocolo, além de contar com editais públicos para fomento de projetos de digitalização e disponibilização de acervos (Ferreira, 2009, p. 10).

É importante notar que a possibilidade de integração dos metadados disponibilizados dessas bibliotecas digitais permite a pesquisadores e pessoas interessadas acesso a grandes bancos de dados para diversas análises da produção cultural. Dependendo da abrangência e da distribuição dessas bibliotecas, podemos ainda considerar a hipótese de analisar toda ou, ao menos, a maioria da produção cultural de uma determinada área do conhecimento, considerando que seus principais documentos estejam presentes  nas bibliotecas digitais de acesso aberto.

Veremos a seguir como o movimento OAI se desenvolveu, seus princípios técnicos, organizacionais e como eles fundamentam um sistema federado de informação, permitindo a criação de ambientes federados.

## 2. Contexto e história do movimento Open Archives Initiative (OAI)

O início do movimento OAI ocorreu num encontro em 1999, em Santa Fe, nos Estados Unidos, conhecido como Convenção de Santa Fe, com o objetivo de discutir como as novas possibilidades de interconectividade da Internet poderiam ser utilizadas para distintas formas de disponibilização de informação acadêmica em rede. O encontro resultou na formação de um grupo de trabalho chamado Open Archives Initiative que tinha por objetivo desenvolver um modelo para facilitar a agregação de provedores de conteúdo na Web (Van de Sompel e Lagoze, 2000).

A Internet e, sobretudo na época, a Web vinha se constituindo como um modelo de comunicação e compartilhamento de conteúdo em rede. O aumento expressivo e a consolidação de padrões como a linguagem HTML[6] e o protocolo HTTP[7] estavam sendo utilizados como base para

4 Portal Scielo:http://www.scielo.br/
5 IBICT – Instituto Brasileiro de Informação em Científica e Tecnológica
6 HTML – HyperText Markup Language
7 HTTP – HyperText Transfer Protocol

inovações na área de sistemas de informação e novos modelos de comunicação.

No entanto, não era apenas a disponibilidade de uma base técnica de comunicação mais eficiente, disponível pela Internet e a World Wide Web, que motivou o encontro de Santa Fe e incentivou a criação da OAI. O paradigma de comunicação científica tradicional também passava por um momento de crise e enfrentava diversos desafios, um consenso que motivava o encontro entre os pesquisadores reunidos em Santa Fe (Lagoze e Van de Sompel, 2001):

- o explosivo crescimento da Internet deu aos pesquisadores acesso universal a um meio de comunicação que facilita o compartilhamento imediato de resultados;
- a rapidez no avanço em muitas áreas do conhecimento tem tornado o modelo de revisão tradicional um impeditivo para o compartilhamento entre pesquisadores;
- a transferência total dos direitos autorais para as editoras das revistas científicas frequentemente age como um impeditivo para o pesquisador que possui interesse em disseminar da forma mais ampla possível seu trabalho;
- a atual implementação do sistema de revisão por pares - um elemento fundamental da comunicação científica – é muito rígido e muitas vezes age sufocando novas ideias, favorecendo artigos de instituições de mais prestígio, causando atrasos para outras publicações;
- o desequilíbrio entre o aumento gradativo dos preços de assinatura de revistas especializadas e o congelamento dos orçamentos das bibliotecas tem criado uma crise econômica para as bibliotecas de pesquisa.

Logo, um dos principais motivos do encontro era reunir pesquisadores e organizações que tivessem interesse em discutir e ajudar a propor as bases técnicas e organizacionais de um novo mecanismo de comunicação científica que buscasse ser uma alternativa aos desafios acima apresentados.

A experiência dos repositórios de *preprints*[8], tendo na iniciativa arXiv.org[9] do Laboratório Nacional de Los Alamos, Estados Unidos, uma das mais conhecidas e disseminadas com foco na área da Física, Ciência da Computação, Matemática e Ciências Não-Lineares, foi extremamente importante para o encontro. Experiências com *preprint* existiam desde 1991, iniciando com o arXiv.org, sendo que muitos outros destes sites estavam se tornando veículos para a disseminação

8 Repositórios de pre-print: repositórios de publicações científicas com possibilidades de envio e acesso a publicações de forma aberta pela Internet sem intermediários. Pode ter ou não recursos de revisão pelos pares dos conteúdos postados.

9 www.arxiv.org

preliminar de resultados de pesquisas e literatura cinzenta[10]. Estes sites acabaram se tornando meios essenciais para compartilhar resultados entre pesquisadores de um campo do conhecimento (Lagoze e Van de Sompel, 2001), dando maior agilidade e facilitando o acesso dos pesquisadores a novos resultados, bem como a aqueles que os produziram. Os repositórios de *preprints* representavam, na época, uma inovação do ponto de vista de como organizar e administrar o conhecimento de organizações, bem como uma nova forma de comunicação acadêmica entre os pesquisadores dessas instituições (Cole e Foulonneau, 2007, p. 47).

A questão que se colocava era como aproveitar a experiência dos repositórios de *preprint* e dar um passo a mais, propondo um modelo de interoperabilidade entre esses repositórios, permitindo que pudessem conversar entre si, compartilhando conteúdos, além da possibilidade de agregação de suas publicações, gerando novos serviços e inovações na área da comunicação científica. Logo, o tema central desse primeiro encontro foi o estabelecimento de recomendações e mecanismos para facilitar o desenvolvimento de serviços entre repositórios de conteúdos. Para que tais recomendações e mecanismos pudessem ser sugeridos, acordos entre os participantes sobre o conceito de interoperabilidade se faziam necessários (Lagoze e Van de Sompel, 2001).

Interoperabilidade tem vários aspectos incluindo uniformização de nomes, formatos de metadados, modelos de documentos, arquitetura de informação, protocolos de acesso, abertura para criação de serviços de terceiros, integração com os mecanismos estabelecidos da comunicação científica, usabilidade entre vários campos do conhecimento, habilidade para contribuir na análise de métricas de uso e citações etc. (Lagoze e Van de Sompel, 2001). Cada um desses aspectos pode ter diferentes interpretações, levando a muitas soluções técnicas possíveis de implementação. O trabalho inicial da OAI era avaliar sugestões e facilitar as negociações entre os membros e comunidades participantes da iniciativa.

Dependendo da forma como a OAI entendesse a interoperabilidade entre sistemas e conforme desenvolvesse suas sugestões de padrões a serem adotados, haveria um impacto direto na possibilidade de ampla adoção ou não dos padrões pela comunidades de potenciais usuários. Havia clareza entre os participantes dos grupos de trabalho da OAI de que criar padrões muito rígidos e especificações muito detalhadas geraria maior resistência e dificuldade para os usuários. Com essa questão em mente, o comitê técnico da OAI entendeu que as recomendações a respeito de interoperabilidade deveriam se restringir ao nível do transporte e compartilhamento de metadados (Lagoze e Van de Sompel, 2001), influenciando e regulando apenas o que o comitê considerou fundamental para que diferentes sistemas de informação pudessem trocar dados.

10 Literatura cinzenta: literatura não convencional, incluindo relatórios, patentes, monografias, teses, dissertações que não foram disponibilizadas por algum meio comercial.

Decisão fundamental, pois permitiu que a iniciativa focasse apenas nas questões relacionadas ao movimento dos metadados em rede, permitindo que outras questões relacionadas a padrões de desenvolvimento de sistemas de informação, as funcionalidades dos sistemas, as estruturas de bancos de dados, interface com usuário, entre outras, pudessem ficar a critério de cada comunidade de desenvolvedores. A escolha da OAI permitiu que diversas soluções de sistemas de repositórios de conteúdos e bibliotecas digitais pudessem atuar com enfoques diferentes, para resolverem demandas diferentes, mantendo, entretanto, o mesmo protocolo para transporte e compartilhamento de metadados, o que tornava os sistemas interoperáveis nesse sentido. A partir dessa escolha, várias aplicações surgiram implementando o padrão OAI, sendo as mais conhecidas citadas a seguir: ePrints[11], Dspace[12], Greenstone[13], FEDORA[14].

Com esse enfoque e a partir da motivação de promover inovações no fluxo de comunicação científica, foi produzida a declaração de missão do movimento OAI, resultado do encontro de Santa Fe, definindo seu escopo e foco de trabalho:

> "A OAI desenvolve e promove padrões de interoperabilidade que objetivam facilitar a disseminação eficiente de conteúdo. A OAI tem suas raízes num esforço para estimular o acesso a repositórios *preprint* como um meio de ampliar a disponibilização da comunicação científica. Contínuo suporte a esse trabalho é a base do programa OAI. Os fundamentos tecnológicos e padrões que são desenvolvidos para suportar esse trabalho são, entretanto, independentes do tipo de conteúdo e dos mecanismos econômicos que regulamentam o conteúdo, se tornando, portanto, promissores na abertura ao acesso de vários outros tipos de objetos digitais, além da comunicação científica. Como resultado, o movimento OAI é uma organização e um esforço explícito na transição e no comprometimento de viabilizar este novo e amplo escopo de aplicações. Conforme ganharmos conhecimento sobre o escopo de aplicação da tecnologia e dos padrões sendo desenvolvidos, e começarmos a compreender a estrutura e a cultura das várias comunidades que irão adotar OAI, nós esperamos realizar contínuas mudanças na missão e na organização da OAI." (Van de Sompel e Lagoze, 2000)

11 www.eprints.org/software
12 www.dspace.org
13 www.greenstone.org
14 www.fedora.info

A missão da OAI evidencia importantes aspectos que possuem relação direta com o modelo de comunicação que pode ser construído utilizando suas especificações. O fato de ter se originado e continuar diretamente conectada com a ideia dos repositórios *preprints* mantém a iniciativa voltada para produzir especificações técnicas e modelos de comunicação que sirvam como alternativas ao modelo tradicional da comunicação científica. A independência do tipo de conteúdo disponibilizado, ou seja, dos padrões e formatos de documentos que podem utilizar essa especificação, cria uma abertura essencial para que o uso dos padrões OAI possa escalar para novos formatos de arquivos e mídias digitais, bem como para outras áreas além da comunicação científica, como a cultura. A independência do modelo econômico posiciona os esforços da iniciativa com foco nos modelos técnicos de comunicação, permitindo que possam ser utilizados com diferentes enfoques, sejam proprietários ou livres. Decisão importante, pois garante que diferentes instituições e comunidades de pesquisadores possam, mesmo com fins econômicos diferentes, serem interoperáveis no nível dos metadados de suas publicações.

Os aspectos organizacionais e culturais explicitados na missão OAI garantem sua intenção de que, por trás de especificações técnicas, possa ser produzido um trabalho contínuo de articulação entre as comunidades usuárias das especificações, mapeando tendências, dificuldades, limitações e possíveis inovações que possam ser incorporadas a iniciativa. O protocolo OAI-PMH, principal especificação da OAI, se encontra em sua versão 2.0, publicada em 14 de junho de 2002, considerada como versão estável pela comunidade para uso no desenvolvimento de aplicações em ambientes de produção.

Analisando os principais elementos que facilitaram a ampla adoção e sucesso na implementação do protocolo OAI-PMH, Carl Lagoze e Herbert Van de Sompel, presidentes do comitê técnico da OAI, consideram as decisões a seguir tomadas pelo comitê como fundamentais (Cole e Foulonneau, 2007, p. 35):

- o modelo organizacional que equilibra liderança efetiva com participação da comunidade: abertura, desde o início, a participação de vários agentes interessados nas especificações, garantindo que múltiplas visões e demandas fossem contempladas ou, ao menos, negociadas. Além de um modelo com intensa participação comunitária, o equilíbrio da liderança efetiva do comitê facilitava o encaminhamento de decisões e agenciamento de impasses que surgiram ao longo do caminho;
- um foco restrito e bem definido do problema a ser resolvido: apesar de inúmeras demandas e sugestões fossem constantemente feitas para serem incorporadas a OAI, o projeto manteve seu foco restrito no nível do transporte e compartilhamento de metadados, garantindo que tudo que estivesse fora desse escopo poderia ser livremente desenvolvido pelas

comunidades que adotassem o padrão;

- um esforço proativo da comunidade: envolvimento intenso e proativo da comunidade, na revisão de especificações, nos testes pilotos e na documentação de problemas;
- um esforço consistente para tomar decisões técnicas transparentes e efetivas através de todo o processo: publicização do modelo de forma aberta e transparente desde o início, garantindo encontros, reuniões, eventos que pudessem definir pontos críticos das especificações.

Ao que parece, uma mistura de bons fatores técnicos e organizacionais emprestou equilíbrio e ritmo a iniciativa, facilitando que soluções simples e inovadoras pudessem surgir como respostas as questões que a OAI se propunha a responder. Vale ressaltar aqui que para futuros desenvolvimento na política de acervos digitais alguns desses elementos mencionados acima deveriam ser considerados, pois instauram um espírito de governança colaborativa que dialoga com os cuidados e investimentos de mediação necessários para o desenvolvimento de uma política com esse enfoque.

Vejamos como o contexto do seu desenvolvimento e as especificações da OAI se constituíram num novo modelo de comunicação em rede para a comunidade científica.

## 3. OAI: interoperabilidade e modelo de comunicação

A proposta da comunidade OAI, em seu contexto organizacional e tecnológico, representa a maneira como a comunidade científica vem utilizando a tecnologia para produzir, disseminar e usar literatura científica estruturada em rede (Weitzel, 2006, p. 87). Oriunda diretamente de uma demanda de melhores estruturas e fluxos de comunicação entre pesquisadores, seu foco se tornou facilitar a disseminação da informação, a busca e o encontro de informação relevante, bem como incentivar a colaboração científica através de um modelo de comunicação que facilitasse a qualquer pesquisador acompanhar o que outros pesquisadores, instituições e centros de pesquisa tem produzido de relevante em sua área de interesse. É a partir dessa perspectiva que se pode entender os repositórios digitais como ferramentas para a promoção da comunicação científica, movimento que origina e inspira a criação da OAI (Bufrem, Gabriel Jr., Gonçalves, 2010).

O ponto chave desse modelo de comunicação proposto pela OAI é a questão da interoperabilidade entre repositórios de conteúdos digitais. Uma das razões para o lançamento da OAI é a crença de que a interoperabilidade entre repositórios é chave para o aumento do seu impacto e no seu estabelecimento como uma alternativa viável ao modelo existente de

comunicação. As vantagens da interoperabilidade podem estimular o uso dos repositórios digitais nos blocos de construção de uma transformação no modelo de comunicação científica (Lagoze e Van de Sompel, 2001).

Um dos objetivos de um modelo de comunicação científica é garantir a mais ampla possibilidade de troca entre os pesquisadores. Considerando que a Internet e a World Wide Web se tornaram um espaço fundamental para a comunicação em rede, essa mais ampla possibilidade de troca entre pesquisadores passa pela capacidade de interoperabilidade entre seus sistemas de informação escolhidos para a publicação do resultado de suas pesquisas.

O modelo OAI foca a questão da interoperabilidade no transporte e no compartilhamento de metadados. Entende-se essa interoperabilidade como a possibilidade de diferentes sistemas de informação publicarem informações sobre suas publicações armazenadas seguindo os mesmos princípios, normas e padrões. Sendo assim, torna-se possível agregar essas informações publicadas e, a partir daí, gerar novos serviços e inovações no uso e processamento da informação.

Novos serviços podem incluir diferentes usos da publicação científica agregada, gerando indicadores, mapas, gráficos, análises bibliométricas e relacionais, bem como novos serviços de busca, monitoramento, acompanhamento de áreas, temas e focos de interesse. A interoperabilidade encoraja a construção de novos serviços (Van de Sompel e Lagoze, 2000), além de ser uma condição fundamental para qualquer modelo de comunicação que pretenda agregar diferentes sistemas de informação distribuídos em rede. É interessante notar como esse interesse que se origina do campo da produção científica dialoga com necessidades muito próximas do que é o campo da produção cultural, sobretudo quando entendemos que o gênero do trabalho que permite essa comparação é praticamente o mesmo, ou seja, a produção oriunda do trabalho intelectual e imaterial.

O protocolo OAI-PMH atua no ponto central deste modelo, viabilizando tecnicamente a circulação da informação em rede. É esse ponto central que viabiliza inovações, como a adoção uma visão federada de sistemas de informação para comunicação científica.

# 4. Sistemas federados de informação

Os sistemas federados de informação surgiram da necessidade de integração de sistemas de informação distribuídos em rede, como uma solução para minimizar dispersão de fontes de dados, reduzir a divergência entre interfaces de busca e ampliar as possibilidades de integração de conteúdos.

Existem várias alternativas de como essa integração pode ocorrer, no entanto, essas alternativas podem ser agrupadas em dois grandes grupos (Marcondes e Sayão, 2001):

- busca distribuída a diferentes servidores: a pergunta de busca é enviada a diferentes servidores, sendo os resultados agrupados e apresentados em uma interface única ao usuário do sistema;
- busca em uma base de metadados centralizada: o sistema realiza um harvesting[15] periódico nos servidores de dados distribuídos, formando um repositório global de metadados. As pesquisas são realizadas nesse repositório global, sendo os usuários redirecionados ao servidor específico de um determinado recurso quando do acesso ao seu conteúdo.

A busca em diferentes servidores é recomendada em situações onde há poucos integrantes e com grandes coleções de dados, do contrário problemas de escalabilidade poderiam ocorrer. Já a busca em uma base de metadados centralizada é recomendada em situações onde a rede é composta de muitos sistemas e se deseja maior agilidade no acesso aos conteúdos, centralizando o processo de busca. As duas soluções acabam por considerar um uso diferente da infra-estrutura de rede, permitindo uma maior ou menor centralização de recursos conforme a demanda e características do tipo de integração que se deseja realizar.

O sistema de base de metadados centralizada, operando através do mecanismo de harvesting, ganhou maior ênfase para o movimento OAI, mostrando-se a solução mais viável para a formação de redes envolvendo vários repositórios digitais (Ferreira e Souto, 2006).

A coleta de metadados vem se tornando, com o movimento OAI, um padrão de organização das redes de bibliotecas digitais, tornando-se um paradigma de ambiente federado de informação. A forma como sua arquitetura de informação foi projetada influencia os aspectos técnicos e organizacionais de como essa rede deve ser estruturada. A arquitetura é baseada na existência de provedores de dados e provedores de serviços, como podemos visualizar na figura 1.

15 Harvesting: sistema de coleta de metadados

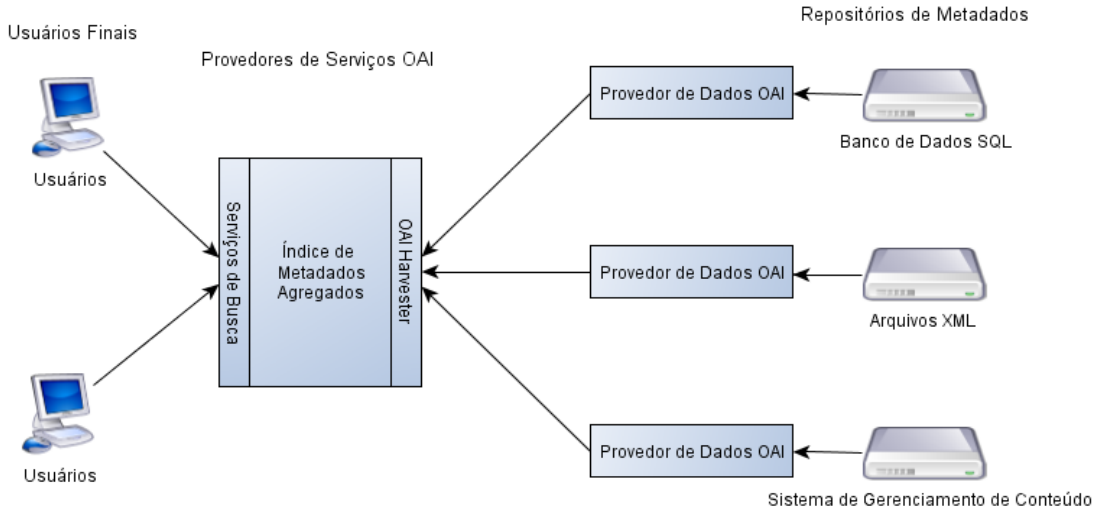



Fig. 1 – Arquitetura de informação OAI

Vejamos como a especificação OAI propõe a definição desses provedores.

## 4.1 Provedores de dados

O provedor de dados é um site na Internet que disponibiliza os conteúdos a serem compartilhados em rede. A especificação OAI não normatiza um formato específico nos quais os documentos de origem precisam ser construídos. Em tese, organizações que possuam uma rotina de disponibilização de conteúdos na Web e um fluxo de produção de metadados poderiam agregar o protocolo OAI-PMH em suas aplicações, tornando-se compatíveis e interoperáveis numa federação de sistemas de informação.

Ao longo dos anos, as comunidades que adotaram o padrão OAI acabaram por adotar formas específicas de como implementar provedores de dados. Os cenários mais comuns são apresentados a seguir (Cole e Foulonneau, 2007, p. 93):

- metadados armazenados como arquivos XML: a arquitetura do repositório assume que os ítens de metadados serão armazenados como arquivos no próprio servidor do sistema de arquivos disponível na Internet. Normalmente, alocam os recursos a serem disponibilizados numa área do servidor e os arquivos de metadados que descrevem esses recursos em outra área;
- metadados armazenados como um banco de dados relacional: é a arquitetura mais utilizada, sendo os metadados armazenados em um banco de dados relacional, normalmente o mesmo banco de dados que armazena os recursos descritos pelos metadados;

- metadados armazenados dentro do recurso que descreve: os próprios arquivos dos recursos a serem disponibilizados incluem uma área alocada para os metadados que descrevem seus recursos
- repositório OAI estático: é representado por um único arquivo XML que contém todos os metadados que são descritos pelo repositório. Normalmente, é utilizado para pequenos repositórios que não são alterados com frequência.

As diferentes formas que podem ser implementadas como um provedor de dados seguindo os padrões OAI dão uma ideia da diversidade e da flexibilidade de soluções que podem ser desenvolvidas. A especificação acaba por atender diferentes necessidades e escalas de repositórios digitais, oferecendo soluções simples e mais complexas, sendo que todas elas podem ser interoperáveis quando agregadas num mesmo provedor de serviços.

## 4.2 Provedores de serviços

O provedor de serviços é um site que realiza o processo de harvesting dos metadados dos sistemas integrantes da federação. É nele que ocorre efetivamente o processo de agregação dos metadados, construindo o índice de metadados agregados, conforme apresentado na figura 1.

O protocolo OAI-PMH não regula os tipos de serviços que podem ser construídos, ficando a critério das comunidades interessadas a produção de diferentes formas de processar, analisar e ofertar possibilidades de interesse geral de manipulação de seus metadados coletados.

Os provedores de serviços acabam se tornando grandes bases de dados que agregam os metadados disponibilizados por todos os sistemas que compõem uma federação. No entanto, a partir da arquitetura da informação a que servem, tornam-se um ponto único de acesso para buscas e análises de todos os repositórios disponíveis nessa rede, verdadeiros pontos de articulação, acesso e disseminação de todo o conteúdo compartilhado em rede. É uma maneira de construir e articular uma rede, não apenas de sistemas de informação, mas de pessoas e organizações que se apropriam desse meio para se comunicarem e compartilharem recursos de interesse.

Logo, esses provedores de serviços tornam-se pontos fundamentais numa federação e espaços privilegiados quando desejamos estudar tendências de como as relações sociais ocorrem entre os membros participantes de uma rede, considerando que esses membros consideram importante ocupar um espaço nessa rede, estar em rede e articular a própria rede em torno de seus objetivos e interesses.

A ideia de utilizar ambientes federados de informação gerou o que a comunidade OAI

passou a chamar de federação de bibliotecas digitais.

## 5. Federação de Bibliotecas Digitais

O conceito de federação surge da necessidade de se denominar processos de integração de diferentes bases de dados e sistemas distribuídos em rede. No caso das bibliotecas digitais, o termo federação aparece quando estamos nos referindo não mais a bibliotecas independentes, mas sim a sistemas que façam a integração de dados de diferentes bibliotecas distribuídas. A federação surge, portanto, como uma denominação a um novo suporte tecnológico que possibilite essa integração de dados (Marcondes e Sayão, 2001). Uma metáfora importante, um conceito que nos permitirá entender uma federação de bibliotecas como uma forma de representar uma rede de bibliotecas digitais.

Vejamos alguns exemplos de iniciativas de federações de bibliotecas digitais que nos auxiliam a perceber como o conceito de federação reflete a maneira como as bibliotecas se organizam em rede:

- Biblioteca Digital de Teses e Dissertações (BDTD)[16]: é um projeto que visa a integração de diversas bibliotecas digitais de Instituições de Ensino Superior brasileiras que disponibilizam teses e dissertações.
- Networked Digital Library of Theses and Dissertations (NDLTD)[17]: é um projeto que visa a integração de bibliotecas digitais de teses e dissertações de vários países do mundo. Foi a primeira experiência de uma federação.
- Networked Computer Science Technical Reference Library (NCSTRL)[18]: é um projeto voltado para a agregação da produção científica da área de Ciências da Computação.
- Univerciencia.org[19]: é um projeto brasileiro que visa a agregação de diversas bibliotecas digitais que disponibilizam recursos voltados para a área das Ciências da Comunicação.

Os exemplos acima apresentados permitem entendermos que uma federação de bibliotecas digitais normalmente representa uma rede temática ou de instituições com fins semelhantes.

A ideia de federação traz a visão da integração de sistemas distribuídos em rede, no caso das bibliotecas, sistemas de bibliotecas digitais. Dentro desse contexto, podemos entender uma biblioteca digital como sendo o nó de uma vasta rede, formando rede a partir de sua

16 www.ibict.br
17 www.ndltd.org
18 www.ncstrl.org
19 www.univerciencia.org

compatibilidade, de sua coerência, de sua padronização com outras inscrições, cada uma das quais representando uma conexão com o mundo através de uma rede (Latour, 2004, p. 49). Esse nó é visto como uma coleção de objetos digitais, juntamente com métodos para acesso e busca da informação, bem como seleção, organização e manutenção da coleção (Witten, Bainbridge e Nichols, 2010, p. 7).

A biblioteca digital torna-se uma presença em rede que representa um contexto, o escopo de representatividade de sua coleção, de sua organização, dos grupos e pesquisadores que a utilizam como uma interface para publicação de seus resultados. O contexto da biblioteca forma um tipo de presença, de integridade conceitual da biblioteca, que empresta coesão e identidade, elementos que refletem sua visão e organização social.

Torna-se, portanto, a presença digital de todo um ciclo de comunicação, representando processos de pesquisa, colaboração e focos de interesse da produção ali depositada. Uma biblioteca digital é um elemento que representa toda uma rede de relações sociais constituídas com o objetivo de fazer cultura, ciência, publicar e compartilhar seus resultados, realimentando o ciclo virtuoso da comunicação, facilitando a circulação e apropriação da informação publicada.

As bibliotecas digitais são produtos de um contexto sobreposto por vários elementos que convergiram em sua presença como um nó de uma vasta rede (Sayão, 2009, p. 7) :

- integração e uso das tecnologias de informação e comunicação;
- maior disponibilização de redes de computadores e tecnologias de apresentação;
- barateamento dos meios de armazenamento em massa;
- disponibilidade crescente de conteúdos digitais em escala planetária;
- possibilidade de digitalização a um custo economicamente viável de conteúdos em mídias convencionais;
- coerência das mídias digitais, abrindo a possibilidade singular para a concepção de novos serviços de informação a partir da integração de objetos digitais.

A biblioteca digital pode ser vista, portanto, como um expoente das novas tendências de sistemas de informação que operam no ambiente distribuído em rede da Internet, aproveitando suas características para viabilizar experiências e novos paradigmas na área da comunicação.

É interessante notar que um dos elementos que caracterizam seu surgimento é a possibilidade de integração de objetos digitais. A digitalização de acervos surge dentro do contexto da interoperabilidade de sistemas, já apontando tendências de produção de novas camadas de funcionalidades, bem como novos modos de organização, sistematização, circulação e apropriação da informação em rede. A biblioteca reproduz no ambiente digital seu objetivo físico, ser um espaço

que promova a organização de acervos ampliando seu acesso e disseminação, sem deixar de lado sua função como espaço de encontro e agregação de produções temáticas.

# 6. Protocolo OAI-PMH

O protocolo OAI-PMH é um modelo de arquitetura de rede cliente-servidor que tem por objetivo regular tecnicamente como deve ocorrer o movimento dos metadados entre um provedor de dados e provedor de serviços, no contexto de um sistema federado de informações. De maneira a facilitar a adoção do protocolo, ele foi todo embasado em vários padrões tecnológicos de comunicação e infra-estrutura em rede amplamente aceitos (Cole e Foulonneau, 2007, p. 21).

O foco de interoperabilidade do protocolo, como mencionamos anteriormente, é o transporte e compartilhamento de metadados. Para que o transporte de metadados possa ocorrer, o protocolo OAI-PMH utiliza três camadas de padrões tecnológicos (ver fig. 2) previamente existentes como infra-estrutura de base em cima da qual seus padrões são construídos: camada de transporte dos dados, camada de linguagem de descrição dos dados e camada de semântica da descrição dos dados.

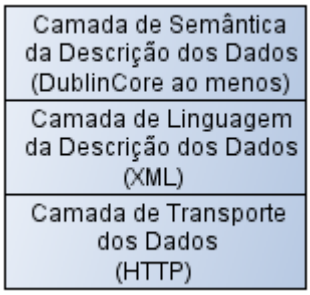



Fig..2 – Camadas de padrões tecnológicos de base para o protocolo OAI-PMH.

As soluções propostas pela comunidade OAI, visando facilitar a adoção do padrão, foram baseadas nos melhores e mais amplamente aceitos padrões a disposição, evitando propor novos padrões para essas camadas. Para o transporte de dados entre provedores de dados e provedores de serviços, a escolha foi utilizar o HTTP[20], procolo baseado na arquitetura cliente-servidor e que serve de base para a Web. Para a linguagem que padronizaria a forma como os metadados deveriam ser descritos, foi escolhido o padrão XML[21], que havia se tornado uma recomendação oficial do World Wide Web Consortium[22] (W3C) em fevereiro de 1998. Para a semântica de descrição dos metadados, o grupo técnico da OAI entendeu que seria necessário propor ao menos um padrão ao qual todos os provedores de dados deveriam respeitar, sendo o padrão escolhido o Dublin Core[23].

20 HTTP – Hypertext Transfer Protocol
21 XML – eXtended Markup Language
22 http://www.w3.org
23 http://dublincore.org

Essa escolha se deve, principalmente, ao entendimento do grupo de que incluir no protocolo OAI-PMH um padrão semântico de descrição de metadados simples seria um elemento facilitador de sua adoção. Vale mencionar que o protocolo OAI-PMH permite a utilização de outros padrões semânticos de metadados, além do DublinCore, não havendo restrições em relação a isso, mas exige que ao menos o DublinCore seja ofertado por qualquer provedor de dados que utilize o padrão.

## 6.1 Papel do HTTP

O protocolo HTTP é baseado na arquitetura cliente-servidor e se propõe a ser um modelo simples de comunicação entre dois computadores em rede, no qual a Web é baseada. O protocolo trabalha com a ideia de sessão de comunicação. Uma sessão representa uma troca de mensagens entre cliente e servidor. O cliente emite um pedido de informação (request message) ao servidor, que envia uma mensagem de resposta (response message), encerrando a sessão entre os dois. Do ponto de vista do protocolo HTTP, qualquer nova transação entre cliente e servidor indica uma nova sessão independente de comunicação entre os dois (Cole e Foulonneau, 2007, p. 24).

Para entender como funciona o transporte de metadados na especificação OAI-PMH, é fundamental entendermos os detalhes de funcionamento do protocolo HTTP.

Um pedido de informação feito por um cliente a um servidor é constituída de 3 partes ( ver figura 3):

- request-line: descreve o método HTTP aplicado pelo cliente, o nome do recurso que o cliente deseja obter e a versão do protocolo HTTP que ele está utilizando. O nome do recurso pode ser um arquivo contendo um conteúdo desejado, como no caso da figura 5.3 o arquivo *test.html* , ou um script de dados que terá de ser executado pelo servidor para devolver a resposta ao cliente. Esse recurso é de extrema importância para viabilizar aplicações mais complexas em rede, permitindo que os clientes da aplicação passem parâmetros a serem utilizados em processamentos específicos pelos servidores;
- request headers: informa ao servidor sobre detalhes do pedido enviado, também dando mais detalhes sobre o que e quem está fazendo o pedido;
- request message body: oferece ao servidor informações adicionais sobre o recurso desejado pelo cliente, como vemos no exemplo da figura 3, no caso da identificação do número de um livro e nome de autor. Normalmente, parâmetros de uma consulta realizada através de formulários são enviados nesta parte numa mensagem HTTP.

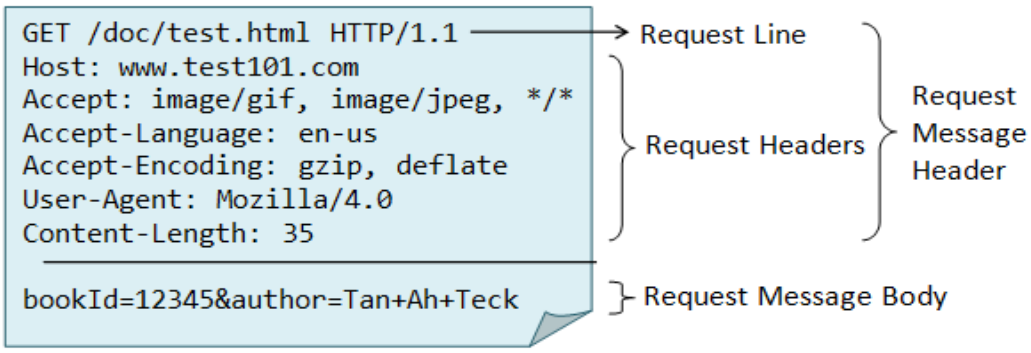



Fig. 3 – Request Message no padrão do protocolo HTTP.

Uma resposta a um pedido de informação também é constituído de 3 partes (ver figura 4):

- response status line: contém uma descrição numérica que representa o estado do servidor, indicando se ele pode ou não responder a requisição ou informando algum erro que possa ter ocorrido;
- response headers: informações sobre o servidor HTTP, sobre o tipo e tamanho de conteúdo que será enviado;
- response message body: representa a resposta enviada pelo servidor, contendo o conteúdo desejado pelo cliente, se o servidor estiver em condições de dispor o recurso. No caso da figura 4, o exemplo apresenta o envio do conteúdo de uma página html, conforme requisitado no exemplo da figura 3.

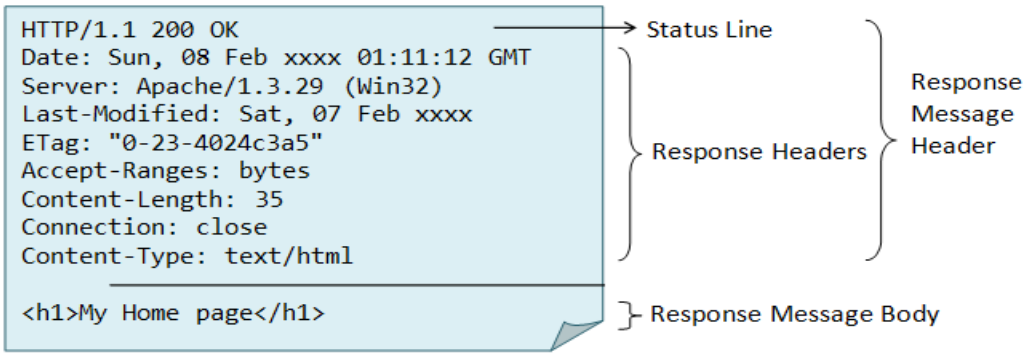



Fig. 4 – Response  Message no padrão do protocolo HTTP.

O protocolo OAI-PMH faz um uso específico do protocolo HTTP, definindo para uma sessão de comunicação o que é um pedido de informação e uma resposta válida entre um provedor de dados e de serviços (Cole e Foulonneau, 2007, p. 25). Dessa forma, cria a possibilidade de sincronização entre os provedores, permitindo que possam se reconhecer como um provedor de dados e um provedor de serviço do padrão OAI-PMH, viabilizando o transporte de metadados entre

eles.

A vantagem de uso do HTTP pelo protocolo OAI-PMH, além de ser um padrão amplamente testado e aceito, é deixar ao seu critério o controle do fluxo de mensagens entre cliente e servidor. O protocolo HTTP utiliza os padrões TCP/IP[24] para o gerenciamento do tamanho das sessões de comunicação, do sequenciamento de pacotes de dados e de como manipular os endereços Web entre cliente e servidor.

## 6.2 Papel do XML

A linguagem XML é uma linguagem de marcação utilizada para definir a estrutura de um documento, originada a partir da linguagem SGML[25], um padrão internacionalmente utilizado desde 1986 como um linguagem para processamento de texto. A estrutura é definida a partir dos elementos que constituem o documento. Elementos são marcados por um início e um fim, delimitados por uma marcação chamada de *tag* (ver fig. 5). Possui algumas regras sintáticas que organizam seus elementos de forma hierárquica, evitando dessa forma que haja sobreposição de elementos entre si.

```
<?xml version="1.0" encoding="UTF-8"?>
<myAlbum>
        <title>Fotografia do Labicom</title>
        <author>
                <familyName>Martins</familyName>
                <givenName>Dalton</givenName>
        </author>
        <publicationDate>24 de agosto de 2015</publicationDate>
        <fileType>PNG</fileType>
        <pages>1</pages>
</myAlbum>
```

Fig. 5 – Exemplo de uma estrutura de um documento XML

Para o uso correto das regras sintáticas da XML, alguns requisitos precisam ser respeitados (Cole e Foulonneau, 2007, p. 28):

- todos os elementos precisam ser explicitamente fechados, mesmo elementos vazios que não possuem nenhum conteúdo;
- todos os atributos de elementos precisam receber um valor e devem ser envolvidos em aspas;
- um elemento XML e o nome de seu atributo deve ser consistente com o conjunto de

24 TCP/IP – Transport Control Protocolo/Internet Protocol
25 SGML – Standard Generalized Markup Language

caracteres utilizados;

- a codificação de caracteres Unicode UTF-8 (8-bit Unicode Transformation Format) por padrão, deve ser mantido ao longo de todo um documento XML.

A XML introduziu três inovações importantes que potencializaram diversas aplicações na Internet baseadas na interoperabilidade entre sistemas de informações, caso do protocolo OAI-PMH. Vejamos quais são:

- estruturas de documentos extensíveis: a semântica da linguagem XML não é fixa, ao contrário da linguagem HTML, por exemplo. Isso permite que qualquer organização ou comunidade de usuários possa criar uma estrutura de um documento padrão e compartilhar entre si as regras semânticas de como esse documento é estruturado. Cada elemento dessa estrutura precisa ter um nome distinto e um significado específico.
- XML Schemas: de forma a explorar melhor a possibilidades de extensão de estruturas de documentos oferecidas, é importante que as aplicações que processarão um documento conheçam as regras semânticas de como ele é organizado. A linguagem XML oferece a possibilidade disso ser informado utilizando um arquivo, chamado XML Schema, que apresenta todas as regras semânticas pelas quais um conjunto de documentos é estruturado. Esse arquivo pode ser utilizado por uma aplicação cliente para entender como processar um conteúdo compartilhado em rede.
- XML Namespaces: um namespace é uma forma de explicitar e manter uma compatibilidade maior da origem das regras semânticas para a construção de um schema XML. Permite que aplicações utilizem um mesmo schema compartilhado entre elas. Um namespace, em termos práticos, é representado por uma URI[26], que é uma forma única de endereçamento de recursos utilizados pela Web, sendo essa maneira de dar um endereço público para um schema XML, podendo ser consultado por qualquer cliente que tenha conhecimento dessa URI. As URIs podem ser criadas por comunidades que desejam declarar qual é seu padrão semântico de metadados, como, por exemplo, a comunidade de pesquisadores de botânica pode publicar uma URI contendo as especificações semânticas de como um arquivo XML deve ser estruturado para seguir seus padrões. Qualquer aplicação interessada em seguir essa convenção, pode utilizar a URI como parâmetro de construção de seus conteúdos XML. A partir deste recurso, é possível combinar diferentes *namespaces* públicos para criar uma convenção particular, mesclando padrões de comunidades.

26 URI – Uniform Resource Identifier

Um documento XML somente é considerado válido quando ele estiver dentro dos padrões de sua estrutura e de acordo com as regras semânticas padronizadas em seu Schema. A possibilidade de validar um documento dessa forma foi uma característica importante a ser considerada pelo protocolo OAI-PMH, que entendeu ser esse recurso fundamental para ofertar um mecanismo de checagem e validação da troca de metadados entre provedores de dados e de serviços. Outro elemento utilizado pelo protocolo OAI-PMH foi o namespace, permitindo que a comunidade utilizasse namespaces já padronizados para compor sua especificação, como é o caso do padrão DublinCore. Essas duas características da XML trouxeram maior flexibilidade, utilização de conjuntos semânticos já padronizados, e robustez, possibilidade de validação da estrutura de um documento, para o protocolo OAI-PMH.

## 6.3 Papel do DublinCore

O padrão Dublin Core é um conjunto de metadados que foi desenvolvido para uso específico por não-especialistas, tendo por intenção ser um conjunto simples e de fácil implementação para a descrição de objetos digitais a serem compartilhados em rede, utilizando a linguagem XML como base. É resultado de um esforço colaborativo de um grande número de comunidades, envolvendo cientistas da computação, bibliotecários e grupos interessados na construções de padrões para descrição de documentos na Web. O primeiro encontro da comunidade Dublin Core ocorreu em Ohio, Estados Unidos, em 1995. O padrão recebeu aprovação do American National Standard Institute (ANSI) em 2001 (Witten, Bainbridge e Nichols, 2010, p. 294).

É formado pelos conjuntos DublinCore Simplicado e DublinCore Qualificado, que é uma forma de ampliar os recursos oferecidos pelo conjunto simplificado, composto de 15 elementos, a serem descritos na tabela 1, a seguir.

| Metadados | Definição |
|---|---|
| Title | O nome dado a um recurso pelo seu criador ou editor |
| Creator | A pessoa ou organização primariamente responsável pelo conteúdo intelectual do recurso |
| Subject | O assunto do recurso |
| Description | Uma descrição textual do conteúdo do recurso |
| Publisher | A entidade responsável por tornar o recurso disponível |
| Contributor | A pessoa ou organização (outra, que não o criador) que é responsável por fazer significativas contribuições ao conteúdo intelectual do recurso |
| Date | Uma data associada com a criação ou disponibilização do recurso |

| | |
|---|---|
| Type | A natureza ou gênero do conteúdo do recurso |
| Format | O formato físico ou digital do recurso |
| Identifier | A referência não-ambígua que pode unicamente identificar o recurso em um dado contexto |
| Source | Uma referência para um segundo recurso de onde o presente recurso possa ter sido derivado |
| Language | A linguagem na qual o conteúdo intelectual do recurso foi desenvolvido |
| Relation | Uma referência a um recurso relacionado e a natureza dessa relação |
| Coverage | Localização espacial e tempo que são características do conteúdo do recurso |
| Rights | Informações sobre os direitos autorais relacionados ao recurso |

Tabela 1. Conjunto de metadados DublinCore Simplificado. Fonte: Witten, Bainbridge e Nichols (2010)

O conjunto qualificado foi criado para facilitar a extensão dos 15 elementos básicos, permitindo que comunidades interessadas em recursos mais expressivos para a descrição de seus conteúdos possam ser contempladas. Qualquer elemento do conjunto simplificado pode ser refinado ou qualificado. Por exemplo, o elemento *date* poderia ser qualificado através dos seguintes elementos *date created, date valid, date available, date issued* ou *date modified.* A característica que torna esse processo de qualificação flexível do padrão DublinCore é o fato de que um elemento qualificado poder ser convertido em simplificado apenas removendo o qualificador, permitindo, dessa forma, que os conjuntos simplificado e qualificado possam co-existir em seus sistemas de informação.

Esquemas de codificação podem regular a faixa de valores permitida para o conteúdo de alguns elementos dos metadados DublinCore. Por exemplo, formato das datas específico, como mm/dd/yyyy, ou um vocabulário controlado para uso no elemento *subject*, regulando a forma como ele deva ser descrito. As restrições oferecidas pelos esquemas de codificação são fundamentais quando pensamos em interoperabilidade de formatos em sistemas distribuídos em rede.

A escolha do DublinCore como referência semântica para os metadados foi uma das decisões mais difíceis e polêmicas tomada pela iniciativa OAI, no entanto, foi baseada na perspectiva da utilização de um padrão estabelecido, simples e de ampla aceitação na Web.

Dentro do contexto de desenvolvimento do protocolo OAI-PMH, a escolha técnica da comunidade OAI para gerar interoperabilidade entre provedor de dados e de serviços foi da utilização do conjunto de metadados DublinCore como padrão semântico mínimo para descrição dos metadados, disponibilizando um XML Schema de descrição do DublinCore Simplificado em conjunto com o protocolo, servindo este para a validação dos dados coletados no sistema de *harvesting*. O *schema* XML é mantido como um *namespace* público, através da publicação de sua URI de acesso, pela Dublin Core Metadata Initiative (DCMI), permitindo que qualquer comunidade interessada possa ter acesso a como esse conjunto de metadados pode ser codificado e decodificado.

Além disso, o transporte de metadados entre os provedores seria feito utilizando o protocolo HTTP. A escolha desses padrões configura o entendimento da comunidade OAI sobre como criar uma arquitetura da informação o mais interoperável possível a partir da utilização de padrões de comunicação e organização em rede já existentes, o que recomendamos que seja feito no âmbito da experiência da política de acervos digitais pelo Ministério da Cultura. Essa recomendação não se baseia apenas na qualidade técnica dos padrões sugeridos mas, sobretudo, no processo social de sua construção, garantindo que muitos atores e em várias etapas tenham feito considerações, debatido contradições e encontrado caminhos de solução para esse modelo de comunicação que deve ser considerado para nossa própria experiência.

# 7. Metadados e normalização

A Internet ganhou muitos adeptos e expandiu de forma expressiva antes que convenções sobre como descrever dados fossem acordadas (Dornfest e Brickley, 2001, p. 205). Fator que indica o rápido processo de apropriação e multiplicidade de serviços que foram desenvolvidos utilizando a rede, mas que também aponta desafios a serem superados quando o que se espera é a integração de sistemas.

O protocolo OAI-PMH fornece toda a estrutura para a construção de ambientes federados de informação. No entanto, além da arquitetura de informação, para que a interoperabilidade dos sistemas possa atingir todo seu potencial, devemos levar em consideração a maneira que uma organização faz a gestão de seus metadados. Estamos falando aqui de aspectos que ultrapassam convenções técnicas e têm relação direta a como essas convenções são apropriadas e utilizadas por quem faz a gestão da informação de uma biblioteca digital. Questões relacionadas a como os metadados são criados, atualizados e eliminados influenciam diretamente a política de coleta que um provedor de serviços deve operar.

Como vimos, os metadados podem ser validados de forma a garantir que se adaptam ao padrão sintático e semântico exigido para o uso em um determinado tipo de serviço. A qualidade e nível de adequação da especificação de metadados que são trocados entre provedor de dados e de serviços influencia diretamente no tipo e na qualidade de serviços que podem ser ofertados. No contexto do protocolo OAI-PMH, os critérios de seleção de metadados podem ser entendidos de 3 formas (Cole e Foulonneau, 2007, p. 139):

- na seleção de qual repositório coletar: indica a possibilidade de escolher qual provedor de serviços será coletado para fazer parte de uma determinada federação;
- como e quando realizar uma coleta seletiva num repositório particular: indica a possibilidade de selecionar um subconjunto de metadados de um determinado

repositório.

- como e quando filtrar os metadados pós-coleta: indica a possibilidade de um provedor de serviços operar diversos procedimentos de filtragem e seleção de metadados conforme a necessidades dos serviços que deseja oferecer.

As duas primeiras formas ocorrem no nível da interação entre provedor de serviços e de dados. A terceira forma ocorre a partir de procedimentos internos que podem ser programados dentro do sistema do provedor de serviços. São procedimentos que atuam diretamente no potencial de agregação dos metadados, buscando melhorar sua qualidade sintática e semântica.

## 7.1 Agregação de Metadados

O processo de agregação pode ser entendido como o conjunto de procedimentos que são necessários para agrupamento dos dados coletados, permitindo que outros procedimentos possam produzir novas informações a partir dessa base comum. São essas novas informações, produto direto da agregação, que podem enriquecer a forma como os documentos compartilhados são apropriados, derivando novos tipos de usos e mecanismos de disseminação.

Provedores de serviços de sucesso, ou seja, aqueles com potencial para atrair maior número de usuários, são aqueles que oferecem serviços avançados de busca, navegação nos dados, suportando buscas a partir de diferentes tipos de entidades, tais como, nomes, títulos, datas, além de serviços de visualização, tais como geração automática de mapas dos repositórios e linhas do tempo (Chavez et al., 2007). Para a efetiva operacionalização desse tipo de serviços, os dados precisam ser tratados em procedimentos internos ao provedor de serviços, após a coleta dos metadados dos provedores de dados.

Vejamos como estes procedimentos podem ser descritos segundo Cole e Foulonneau (2007, p. 155):

| Selecionar | Limpar | Normalizar | Aumentar | Adaptar |
|---|---|---|---|---|
| Excluir registros que não correspondem a política de uso do provedor de serviços. Ex.: dados que possuem direitos autorais divergentes. | Remover elementos de concatenação. Ex.: pontuações, marcadores de começo e fim de uma sentença. | Renomear campos e/ou mapeá-los de um campo para outro. Ex.: um registro vem com o nome do campo Autor e outro Author. | Acrescentar valores e/ou detalhar campos. Atribuir valores-padrão para todos os registros de um mesmo repositório. Ex.: acrescentar nome de instituição a dados provenientes de um mesmo local. | Selecionar os registros que serão utilizados por um determinado serviço. |

| | | | | |
|---|---|---|---|---|
| Remover registros duplicados ou reconciliar metadados que descrevem objetos com a mesma URI. | Remover campos vazios. | Modificar/transformar valores para vocabulários controlados e/ou normalizar os valores. Ex.: padronizar a forma como os registros descrevem o assunto que representam | Acrescentar nome de uma coleção e outros campos e valores que forem pertinentes ao contexto do repositório. | Selecionar campos alternativos quando a primeira opção não estiver disponível. Ex.: não há nome do autor, mas existe o campo sobrenome. |
| | Separar valores que foram concatenados. Ex.: separar em dois campos quando nome e sobrenome vierem juntos. | | Relacionar os dados a uma autoridade externa. Ex.: atribuir aos dados o nome da instituição financiadora do projeto. | Decidir estratégias para quebrar valores e listar múltiplos valores. Ex.: tratamento de um campo data, exibindo apenas dia, ano ou mês. |

Tabela 2. Procedimentos internos de tratamento dos metadados pós-coleta. Fonte: Cole e Foulonneau (2007)

Certamente, nem todos os procedimentos apresentados na tabela 2 precisam ser implementados por um provedor de serviços. Os procedimentos a serem utilizados vão variar em relação a qualidade de produção dos metadados dos provedores de dados, sendo que podem estar mais ou menos alinhados em torno de um mesmo propósito e de normas comuns para a publicação de informação em rede.

A título de ilustração, vale a pena mencionarmos um estudo que avaliou como os itens *subject* e *description* do padrão Dublin Core Simplificado foram utilizados por três tipos diferentes de instituições:

| | **% de registros coletados contendo o elemento** | |
|---|---|---|
| | **Subject** | **Description** |
| **Bibliotecas digitais** | 78 | 36 |
| **Museus e sociedades históricas** | 93 | 93 |
| **Bibliotecas acadêmicas** | 15 | 13 |

Tabela 3. Variação no uso de dois elementos Dublin Core por tipos de instituição. Fonte: Cole e Foulonneau (2007, p. 170)

Um outro estudo (Ward, 2004) parece confirmar a tabela acima, indicando uma grande variação no uso de campos Dublin Core pelos provedores de dados. Analisando 82 provedores de dados e 910.919 registros de metadados no padrão Dublin Core Simplificado, os resultados indicaram que 54% dos provedores utilizavam apenas os campos ***creator*** e ***identifier*** para aproximadamente 50% dos metadados que disponibilizam.

A variação dos dados apresentada acima nos permite concluir que, por exemplo, um serviço

que pretenda fazer uma análise dos assuntos disponibilizados por essas coleções pode operar de forma significativa no contexto dos Museus e Bibliotecas Digitais, tornando-se praticamente irrelevante no contexto das Bibliotecas acadêmicas por falta de dados. O mesmo se passa no quesito descrição dos recursos, inviabilizando desta vez também as Bibliotecas Digitais por falta de informações abrangentes. Dependendo do contexto e dos recursos disponíveis para um projeto, estes campos podem ser complementados pelo provedor de serviços, ou mesmo um acordo entre provedor de serviços e de dados pode levar a melhorias nessas taxas, ampliando o nível de colaboração entre os provedores.

Sendo assim, o desenvolvimento de um novo serviço deve levar em consideração uma análise prévia da qualidade dos metadados de forma que possa projetar quais serão os procedimentos pós-coleta que precisam ser implementados.

## 7.2 Qualidade dos Metadados

A qualidade de um conjunto de metadados deve ser avaliada levando em consideração o propósito pelo qual um repositório foi criado. Os metadados podem atender as necessidades e demandas de um repositório, a partir de seu contexto local. Diversos tipos de problemas locais podem influenciar na qualidade dos metadados, desde erros tipográficos ao processo de conversão de dados para formatos digitais (Beal, 2005). No entanto, no contexto de uma federação a qualidade dos dados de um repositório pode decair em função da necessidade de integração com outros sistemas de informação.

A integração dos sistemas leva a três questões relacionadas diretamente a qualidade dos metadados (Nichols, McKay e Twidale, 2008). Primeiro, múltiplos formatos da semântica de metadados podem estar presentes na federação, levando a necessidade de procedimentos de conversação de um padrão em outro para formar uma coleção única, acarretando perda de informação. Segundo, diferentes projetos, mesmo utilizando o mesmo formato de metadados, podem ter entendimentos distintos de como um campo deva ser preenchido, levando a inconsistências que precisam ser tratadas quando da formação de uma coleção única. Terceiro, um repositório pode assumir um contexto local e desenvolver os seus dados a partir deste contexto não o explicitando em seus metadados. Um exemplo de como isso pode acarretar inconsistências é um repositório dedicado a um evento histórico específico, não descrevendo em seus metadados esse evento, levando em consideração que a busca de informações seria sempre a partir desse contexto. Quando esses dados fossem agregados em uma coleção maior, o contexto do evento teria se perdido.

Logo, essas três questões devem ser levadas em consideração quando da avaliação da

qualidade dos metadados e na especificação dos procedimentos de tratamento. De forma a facilitar uma avaliação sistemática da qualidade de um conjunto de metadados, podemos avaliar seu padrão de qualidade segundo sete dimensões (Witten, Bainbridge e Nichols, 2010, p. 323):

- completude: o nível de campos que se encontram preenchidos de informação. No nosso caso, estamos falando dos campos referentes minimamente ao padrão Dublin Core;
- precisão: a quantidade de erros que pode ser encontrada;
- proveniência: a fonte de onde provém os metadados. No nosso caso, a proveniência nos será útil quando quando forem claramente reconhecidos como oriundos do campo das Ciências da Comunicação;
- ajuste aos padrões: o nível de ajuste as especificações semânticas e sintáticas. No nosso caso, ajuste aos padrões propostos pelo protocolo OAI-PMH;
- consistência lógica e coerência: o nível de consistência dentro de todo o conjunto de um repositório. No nosso caso, diz respeito a maneira e os critérios pelos quais os metadados são preenchidos pela biblioteca digital de origem, no provedor de dados;
- padrão no tempo: o nível periódico de atualização de um repositório. No nosso caso, a clareza quanto ao padrão de datas utilizados nos documentos;
- acessibilidade: a forma como os dados podem ser acessados. No nosso caso, dados disponíveis para serem coletados pelo *harvester.*

Uma questão importante que teremos de analisar é a qualidade, sob os critérios acima, dos metadados que iremos coletar como fonte de serviços informacionais para a área da cultura. Essa é uma questão determinante do nível de profundidade onde podemos chegar em nossas análises, dado que quanto mais os metadados coletados se ajustarem a esses padrões de referência maior será nossa possibilidade de manipulação dos mesmos. Se camadas expressivas de dados estiverem faltando ou operando a partir de consistências lógicas diferentes, por exemplo, teremos de encontrar alternativas para normalizar esses dados utilizando outros sistemas de informação como apoio ou teremos de descartar determinados níveis de análise, devido a possíveis dificuldades técnicas.

# 8. Conclusão

O presente relatório procurou colocar em perspectiva o modelo de comunicação, os padrões técnicos e as práticas de gestão que podem ser integradas para a constituição de um modelo federado de integração de repositórios de documentos digitais. A complexidade das múltiplas camadas que devem ser levadas em consideração, bem como os diferentes níveis de decisão que devem ser tomadas para a execução de um projeto desse porte, considerando a abrangência política

e territorial do Ministério da Cultura, evidenciam que as práticas e instâncias de governança dessa política devem refletir essa complexidade.

A escolha de padrões de abertos e a conexão direta com experiências que possuem alta relevância social são fundamentais para ampliar a chance de êxito de projetos nessa área, dado que se baseiam em práticas e em processos de avaliação técnica testados por muitas comunidades de técnicos e usuários de sistemas de informação, evitando a endogenia de soluções que podem ser facilmente descontinuadas ou se mostrarem frágeis em aspectos fundamentais que não foram considerados em seu momento de adoção.

Entendemos que a arquitetura federada ainda se constitui em um desafio complexo de realização, dado a dificuldade de articulação de todos os aspectos mencionados acima, porém se mostra uma aposta potente e fundamental para o exercício de produção de redes de conteúdos digitais na área da cultura.

## 9. Referências


BEAL, J., Metadata and data quality problems in the Digital Library. **Journal of Digital Information**, Vol 6, No 3. 2005.

BUFREM, L. S., GABRIEL JR., R. F., GONÇALVES, V. Práticas de co-autoria no processo de comunicação científica na pós-graduação em Ciência da Informação no Brasil. **Inf. Inf.,** Londrina, v. 15. n. esp. p. 110-129, 2010.

CHAVEZ, R., CRANE, G., SAUER, A., BABEU, A., PACKEL, A., WEAVER, G. ,Services make the repository. **Journal of Digital Information**. Vol.8, no. 2, 2007.

COLE, T. W., FOULONNEAU, M., **Using Open Archives Initiative Protocol for metadata harvesting**. Libraries Unlimited. 2007. 224p.

DORNFEST, R., BRICKLEY, D. Metadados. In. Oram, A. (org.) **Peer-to-peer: o poder transformador das redes ponto a ponto.** Editora Berkeley. 2001. 447p.

FERREIRA, S. M. S. P., SOUTO, L. F., Dos sistemas de informação federados à federação de bibliotecas digitais. **Revista Brasileira de Biblioteconomia e Documentação**, Nova Série, São Paulo, v. 2, n. 1, p. 23-40, jan/jun. 2006.

FERREIRA, S. M. S. P., **Ferramenta de busca federada de Teses e Dissertações para aplicação em áreas especializadas.** Relatório Técnico. Processo CNPq. no. 480927/2007-3. 2009

LAGOZE, C., VAN DE SOMPEL, H., The Open Archives Initiative: building a low-barrier interoperability framework. **JCDL'01,** June 17-23, 2001.

LATOUR, B. Razão que a razão desconhece: laboratórios, bibliotecas, coleções. In. Parente, A.

(org.), **Tramas da Rede**. Sulina. 2004. 303p.

MARCONDES, C. H., SAYÃO, L. F., Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ci. Inf.** Vol. 30, no.3 Brasília, set/dez 2001.

NICHOLS, D. M., McKAY, D., TWIDALE, M. B., A lightweight metadata quality tool. **JCDL'08,** june 16-20, 2008.

SAYÃO, L. F., Afinal, o que é biblioteca digital? In.: Bibliotecas Digitais/Bibliotecas Virtuais. **Revista USP**. D zembro/janeiro/fevereiro 2008-2009.

VAN DE SOMPEL, H., LAGOZE, C. The Santa Fe Convention of the Open Archives Initiative. **D-Lib Magazine**, vol. 6, no. 2, February, 2000.

WARD, J., Unqualified Dublin Core usage in OAI-PMH data providers. **OCLC Systems and Services: International Digital Libraries Perspectives.** Vol. 20. no. 1. 2004. p. 40-47.

WEITZEL, S. R. Fluxo da comunicação científica. In. Poblacion, D. A., Witter, G.,Silva, J. F. M. (org.). **Comunicação e produção científica: contexto, indicadores,avaliação.** Angellara, São Paulo, 2006, pag. 82 – 114. 428p.

WITTEN, I. H., BAINBRIDGE, D., NICHOLS, D. M. **How to build a digital library.** Morgan Kauffmann Publishers. 2nd edition. 2010. 656p.