Protocolo OAI-PMH e Sistemas Federados de Informação: fundamentos de arquitetura da informação para análise de dados do portal de produção científica da área de Ciências da Comunicação Univerciencia.org

Dalton Lopes Martins[*]

Sueli Mara Soares Pinto Ferreira[**]

Resumo O movimento de Arquivos Abertos através do protocolo OAI-PMH tem facilitado a criação de federações de bibliotecas digitais que permitem ampla agregação de produção científica, criando condições para a construção de importantes bases de dados para pesquisa bibliométrica, cientométrica, análise de redes sociais e outros modos de análise. Como um meio privilegiado para avaliação e identificação de padrões de produção da informação em diferentes comunidades científicas, as bibliotecas digitais federadas podem ser utilizadas como fontes de dados para pesquisa científica. A construção dessas bases de dados envolve diversos passos iniciando pelo modo com que a informação deve ser estruturada pelas bibliotecas locais, a forma como deve ser compartilhada em um ambiente de rede e sintetizada de modo a garantir visualizações integradas da informação. Questões relevantes sobre a qualidade dos metadados também são abordadas do ponto de vista metodológico e do modo como impactam as possibilidades de análise científica. Apresentamos neste artigo os fundamentos de como essa arquitetura de informação foi aplicada na construção do portal Univerciencia.org, com foco específico na produção científica da área da Ciências da Comunicação, bem como a tecnologia utilizada, as formas de estruturação e tratamento dos metadados, além dos primeiros resultados da análise de dados coletados e sua qualidade.

Palavras-chave bibliotecas digitais, sistemas federados, OAI-PMH, repositórios abertos, cientometria, bibliometria.

OAI-PMH protocol and Federated Information Systems: fundaments of information architecture for the analysis of data of the scientific production portal in the Comunication Sciences Univerciencia.org area

Abstract The Open Archives Initiative movement using the OAI-PMH protocol has allowed the production of federated digital libraries that provide wide aggregation of scientific production, creating conditions for the construction of important databases that enable the development of bibliometric research, scientometric, social network analysis and other modes of analysis. As a privileged means for evaluating and identifying patterns of information production in different

[*] Professor doutor na área de redes. Faculdade de Tecnologia de São Paulo (FATEC-SP). Praça Cel. Fernando Prestes, 30 – Edifício Santhiago – 1º andar, Bom Retiro, 01124-060 São Paulo-SP. Telefone: (11) 3322-2218. E-mail: dmartins@gmail.com.

[**] Professora titular e diretora técnica do Sistema Integrado de Bibliotecas [SIBi/USP]. Rua da Praça do Relógio, 109, Bloco K, 4º andar – Prédio da Administração Central, Cidade Universitária, 05508-010 – São Paulo – SP. Telefone: (11) 3091-4195 e (11) 3091-1547. E-mail: sueli.ferreira@gmail.com.

scientific communities, federated digital libraries can be used as a source to study these communities. The construction of these databases involves several steps, beginning with the way information should be structured by local libraries, shared in a network environment and synthesized in order to ensure an integrated view of information. Relevant questions on the quality of such data directly impact on the possibilities of scientific analysis. We present here the foundations of how that information architecture has been implemented in the construction of the portal Univerciencia.org, which is focused on the scientific production of Communication Science, as well as the technologies involved, the ways of structuring metadata and the first results of analyzes of the data collected and their metadata quality.

Keywords digital libraries, federated systems, OAI-PMH open repositories, scientometrics, bibliometrics.

Introdução

O movimento Open Archives Initiative[1] 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. 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.

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), envolvendo diversos tipos de instituições. Dentre elas, estão 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 2012 já temos 95 bibliotecas de teses e dissertações no IBICT[4], muitas delas disponibilizando metadados com protocolo OAI-PMH, além de contar com editais públicos para fomento de projetos de digitalização e disponibilização de acervos (Ferreira, 2009, p. 10). No entanto, ainda há muito para ser feito, considerando as dimensões de um país como o Brasil.

[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

4 IBICT – Instituto Brasileiro de Informação em Científica e Tecnológica

É importante notar que a possibilidade de integração dos metadados disponibilizados dessas bibliotecas digitais permite aos pesquisadores estudarem grandes bancos de dados para diversas análises da produção científica. Dependendo da abrangência e da distribuição dessas bibliotecas, podemos ainda considerar a hipótese de analisar toda ou, ao menos, a maioria da produção científica de uma determinada área do conhecimento, considerando que suas principais instituições e pesquisadores publiquem sua produção em revistas e bibliotecas digitais de teses e dissertações de acesso aberto.

O objetivo deste artigo é demonstrar como a arquitetura da informação proposta pelo protocolo OAI-PMH tem sido utilizada pelo portal Univerciencia.org com foco específico na área da Ciência da Comunicação. Serão apresentados os resultados das primeiras experiências de coleta e síntese dos dados, além dos desafios e benefícios deste tipo de arquitetura para a construção de bases de dados significativas para geração de indicadores e análises da produção científica de uma área do conhecimento. Desenvolve-se uma discussão das diversas etapas da estruturação, compartilhamento e síntese da informação, fornecendo o embasamento sistêmico informacional para a construção de uma biblioteca digital federada.

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 tornou-se 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 facilite a qualquer pesquisador acompanhar o que outros pesquisadores, instituições e centros de pesquisa têm 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 (Bufrem, Gabriel Jr., Gonçalves, 2010).

O ponto chave do modelo proposto pela OAI é a 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 desse 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. A interoperabilidade permite ampla agregação de dados, criando condições de novas sínteses de informação, condição fundamental para uma visão científica que utiliza a lógica indutiva como um importante instrumento de análise, tal como nosso modo atual de fazer ciência.

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, ao gerar 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.

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

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 a 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, mas 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*[5] 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 em que 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 usos diferentes 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*, deu maior ênfase ao movimento OAI, mostrando-se a solução mais viável para a formação de redes

5 Harvesting: sistema de coleta de metadados

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, sendo 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.

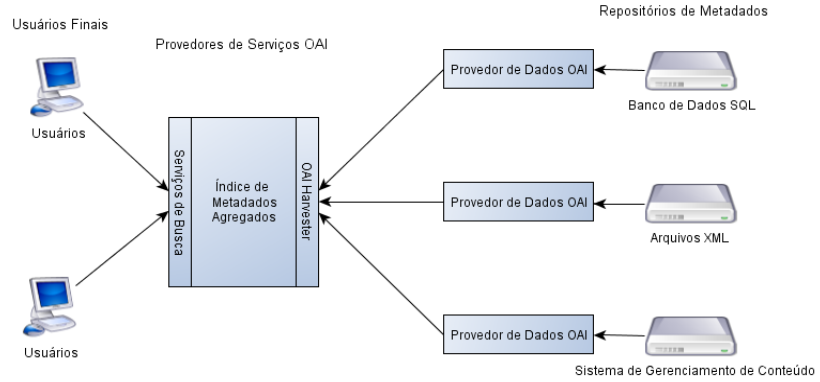

Figura 1: Arquitetura de informação OAI

Tecnologia de base

A construção do portal Univerciencia.org é baseada na aplicação dos princípios do protocolo OAI-PMH, que veremos a seguir. Além disso, nesta seção, veremos questões metodológicas fundamentais no trato dos metadados, que implicam diretamente a qualidade da agregação e análise das informações coletadas pelo protocolo. São esses princípios metodológicos que orientam o processo de coleta e análise posterior dos dados que têm sido utilizados nas pesquisas que temos desenvolvido.

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 figura 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.

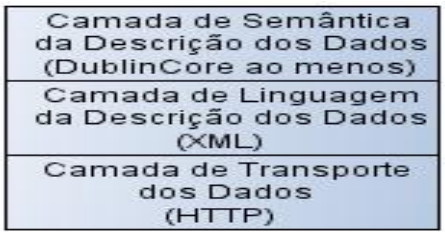

Figura 2: Camadas de padrões tecnológicos de base para o protocolo OAI-PMH.

As soluções propostas pela comunidade OAI, visando facilitar a adoção do padrão, foram baseadas nos melhores e mais amplamente aceitos padrões tecnológicos à 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[6], 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[7], que havia se tornado uma recomendação oficial do World Wide Web Consortium[8] (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 que todos os provedores de dados deveriam respeitar, sendo o padrão escolhido o DublinCore[9]. Essa escolha deve-se, principalmente, ao entendimento do grupo de que incluir no protocolo OAI-PMH um padrão semântico simples de descrição de metadados 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.

Vejamos agora qual o papel dos metadados para a arquitetura de informação do portal.

Metadados

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, devemos levar em consideração a maneira como uma organização faz a gestão de seus metadados para que a interoperabilidade dos sistemas possa atingir todo seu potencial. Estamos falando aqui de aspectos que ultrapassam convenções técnicas e têm relação direta com a maneira 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

6 HTTP – Hypertext Transfer Protocol
7 XML – eXtended Markup Language
8 http://www.w3.org
9 http://dublincore.org

serviços influencia diretamente o tipo e a 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 as 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 metodológicos que atuam diretamente no potencial de agregação dos metadados, buscando melhorar sua qualidade sintática e semântica.

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 1:. 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, podendo 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 DublinCore Simplificado foram utilizados por três tipos diferentes de instituições (Cole e Foulonneau, 2007, p. 170):

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

Tabela 2:. 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 estabelecer um acordo entre provedor de serviços e de dados que 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.

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 a 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 tres questões relacionadas diretamente à qualidade dos metadados (Nichols, McKay e Twidale, 2008). Primeiro, múltiplos formatos da semântica de metadados podem estar presentes na federação, levando à 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: a taxa 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 encontrada;

- proveniência: a fonte de onde provêm os metadados. No nosso caso, a proveniência será útil quando forem claramente reconhecidos como oriundos do campo das Ciências da Comunicação;

- ajuste aos padrões: o nível de ajuste às 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 à maneira e os critérios pelos quais os metadados são preenchidos pela biblioteca digital;

- padrão no tempo: o nível periódico de atualização de um repositório. No nosso caso, a clareza quando ao padrão de datas utilizado 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 análise. 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, como por exemplo o sistema Lattes, ou teremos de descartar determinados níveis de análise, devido a possíveis dificuldades técnicas.

Resultados: o portal da produção científica em Ciências da Comunicação Univerciencia.org

O portal da produção científica em Ciências da Comunicação – Univerciencia.org – é uma iniciativa do CEDUS - Centro de Estudos em Design de Sistemas Virtuais da Escola de Comunicações e Artes da USP, que vem sendo implementado como parte dos projetos e estudos que tiveram início a partir de 2000.

Para cada fonte de informação o portal pode coletar diferentes recursos informacionais. Recorremos ao texto de apresentação do portal para sua definição sobre recurso informacional: reporta-se à tipologia dos conteúdos publicados pelas fontes que estão sendo catalogadas. Até o momento as tipologias coletadas pelo Portal estão identificadas da seguinte forma:

1. anais de eventos;

2. artigos de revistas;

3. comunicações feitas em eventos;

4. livros e capítulos de livros;

5. miscelâneas (cobre relatórios, trabalhos de pesquisa, publicações técnicas dentre

outras. (Usado também para fontes que contêm mais de um tipo de recurso.) ;

6. recursos educacionais abertos (REA);
7. revistas;
8. trabalhos de conclusão de curso (TCC);
9. teses e dissertações.

É importante destacar que serão esses recursos informacionais coletados pelo portal que teremos a nossa disposição para as análises desenvolvidas.

Vejamos agora de que amostra de dados estamos falando, apresentando a seguir as principais características do banco de dados do Univerciencia até o presente momento em que escrevemos.

O portal possui 35.785 recursos informacionais coletados de 17 países diferentes, representando 98 fontes de informação de 68 instituições diferentes. Vejamos como esses dados estão distribuídos na tabela 3.

País	Itens	Fontes	Instituições	Tipo de Recurso	%Itens
Brasil	19235	60	36	9	53,75%
França	3310	1	1	1	9,25%
Espanha	3178	8	6	4	8,88%
México	2808	3	3	2	7,85%
Estados Unidos	1859	4	4	3	5,19%
Canadá	1807	2	1	1	5,05%
Dinamarca	890	1	1	1	2,49%
Portugal	747	8	6	6	2,09%
Equador	467	1	1	1	1,31%
Colômbia	430	2	2	1	1,20%
Suíça	309	1	1	1	0,86%
Costa Rica	290	1	1	1	0,81%
Argentina	205	2	1	2	0,57%
Filipinas	81	1	1	1	0,23%
Venezuela	79	1	1	1	0,22%
Austrália	50	1	1	1	0,14%
Indonésia	40	1	1	1	0,11%
17	**35785**	**98**	**68**	**em 12**	**100,00%**

Tabela 3:. Dados da distribuição dos recursos informacionais por países – Fonte. www.univerciencia.org – Acessado em 25/06/2011.

Vejamos como esses recursos informacionais estão distribuídos nas tipologias de recursos coletados pelo portal na tabela 4.

Tipo de Recurso	Itens	Fontes	Instituições	Países	%Itens
Revista	16925	62	52	15	47,30%
Miscelânea	6432	5	5	4	17,97%
CBCC	3536	1	1	1	9,88%
Dissertação/Tese	3481	17	17	4	9,73%
Livro/Capítulo	2491	2	2	2	6,96%
Artigos	1782	2	2	2	4,98%
Comunicação em Evento	465	2	2	2	1,30%
Anais	248	1	1	1	0,69%
Anuário	237	3	3	3	0,66%
ENDECOM	97	1	1	1	0,27%
COLÓQUIO	89	1	1	1	0,25%
REA	2	1	1	1	0,01%
12	**35785**	**98**	**em 68**	**em 17**	**100,00%**

Tabela 4: Dados da distribuição dos recursos informacionais por tipos de recursos – Fonte. www.univerciencia.org – Acessado em 25/06/2011.

Fica visível a abrangência da base de dados do portal a partir dos dados acima apresentados. Estão à disposição mais de 35.000 recursos informacionais coletados, sendo a sua maior parte proveniente do Brasil e mais de 47% de artigos publicados em revistas científicas da área. Considerando esse papel do Brasil na base de dados e a importância que artigos de revista têm na mesma, analisamos a seguir mais detalhadamente como se dá o comportamento das revistas científicas e seus principais resultados. Nosso objetivo é descrever como esses recursos têm sido apropriados pela produção científica brasileira na área da Ciência da Comunicação.

Apresentamos na figura 3 a distribuição ao longo do tempo das revistas ativas, ou seja, aquelas que forneceram metadados de artigos nos respectivos anos da distribuição.

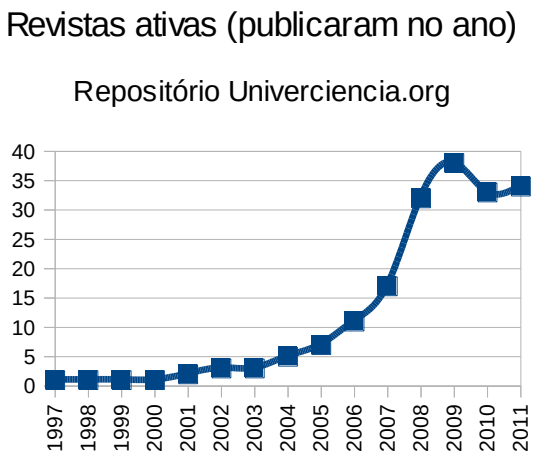

Figura 3: Distribuição das revistas ativas ao longo do tempo. Fonte: www.univerciencia.org

É interessante notarmos o aumento expressivo no número de revistas ativas no portal do ano de 2007 a 2008, levando a crer que ocorreu algum fenômeno específico nesse momento de crescimento no número de revistas convergindo na maior adoção pelo campo da Ciência da

Comunicação do modelo OAI-PMH como modelo de distribuição de seu conteúdo. Podemos relacionar esse aumento com o fato de a CAPES e o CNPq passarem, a partir de 2007, a financiar diretamente a editoração de revistas brasileiras em formato aberto, com investimentos de aproximadamente U$ 2.4 milhões por ano em conjunto[10]. O efeito do fomento de recursos dessa política pública de incentivo à editoração de revistas que é percebido na figura 3, no caso das Ciências da Comunicação através da base Univerciencia.org, pode ser também confirmado analisando as revistas que foram beneficiadas no resultado do edital de 2007[11], onde podemos encontrar as revistas Interface e Ciências & Cognição, estando elas entre as que mais forneceram artigos para o portal, entre outras.

Observando como se distribuem os metadados coletados com enfoque nas revistas científicas brasileiras, obtivemos 49 fontes de artigos que foram coletadas pelo portal, com mais de 9.800 artigos disponíveis para análise. Essas fontes e o número de artigos que cada uma fornece podem ser vistas na figura 4.

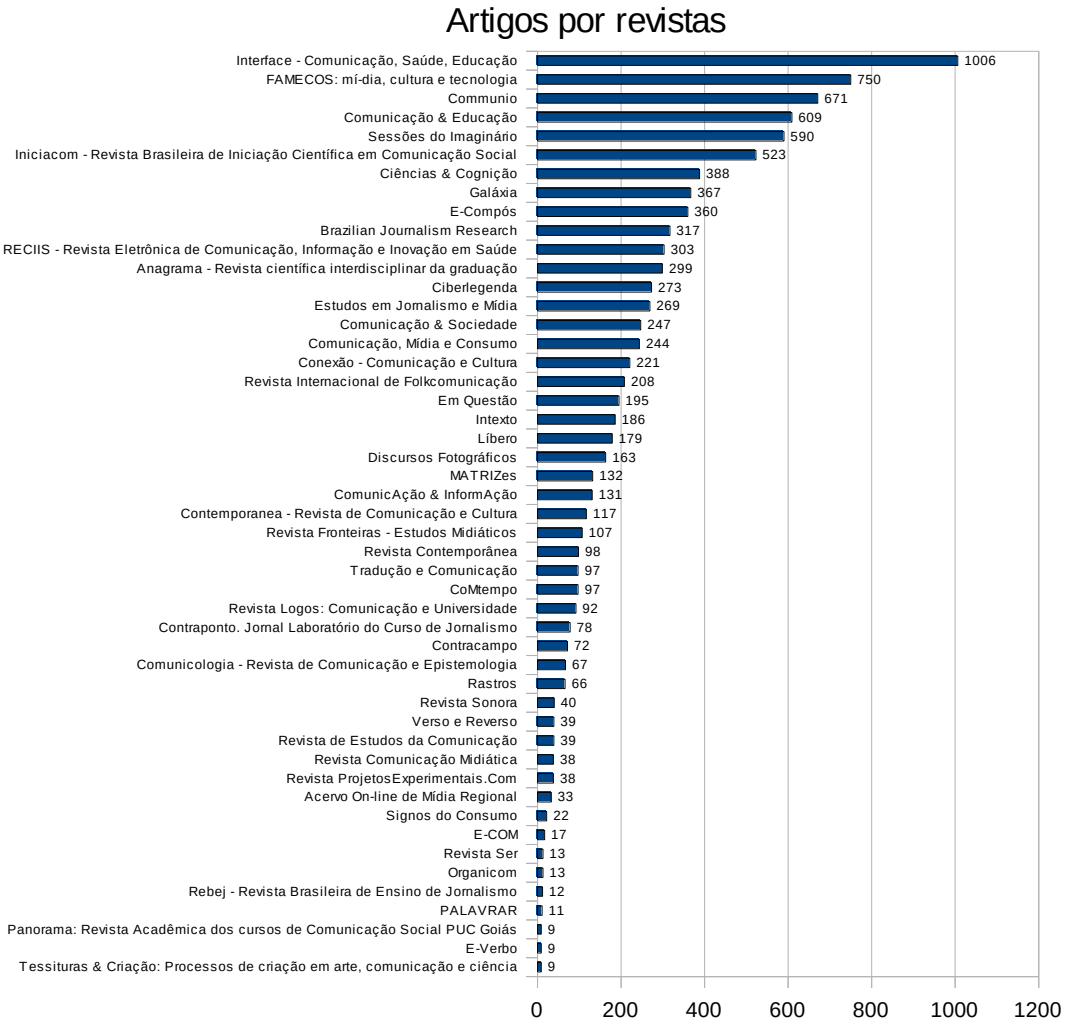

Figura 4: Distribuição dos artigos por revistas científicas. Fonte: www.univerciencia.org

10 http://www.capes.gov.br/servicos/sala-de-imprensa/artigos/4720-as-razoes-para-o-avanco-da-producao-cientifica-brasileira

11 http://www.cnpq.br/resultados/2007/docs/016.pdf

As revistas mais antigas tendem a apresentar uma maior quantidade de artigos, e as mais novas apenas artigos de poucas edições publicadas, demonstrando como se dá a dinâmica de distribuição de artigos publicados entre os periódicos disponíveis no portal.

Para avaliar como ocorria a distribuição do número de coautores por artigo nos metadados coletados no portal, realizamos uma análise do padrão de preenchimento do campo *dc:creator* (o campo responsável por apresentar nos metadados o autor(es) do conteúdo, segundo a tabela 1) em cada um dos registros das revistas brasileiras. Na base de dados do portal constavam 9.864 registros representando o número de artigos coletados das revistas brasileiras. Analisando o resultado do preenchimento desses registros através de um programa que procurava automaticamente o número de autores de cada artigo baseado na quantidade de campos *dc:creator* que constavam em cada registro, bem como se eles seguiam as recomendações semânticas do padrão XML, encontramos 9654 registros válidos segundo nossos critérios. O resultado representa um total de 97,87% dos registros válidos para serem considerados na análise de coautoria. Sem dúvida, o resultado indica para este campo dos metadados um número expressivo, garantindo que as análises feitas com base nos dados coletados possam ser representativas e abrangentes em relação ao que é disponibilizado pelo portal.

A seguir, calculamos a média de coautoria dos artigos agrupados por periódico, para analisar como se dava essa distribuição em relação às 49 fontes de dados utilizadas em nossa análise. A média ponderada apresentou um resultado de 1,4330 autores por artigo. Vejamos essa distribuição na figura 5.

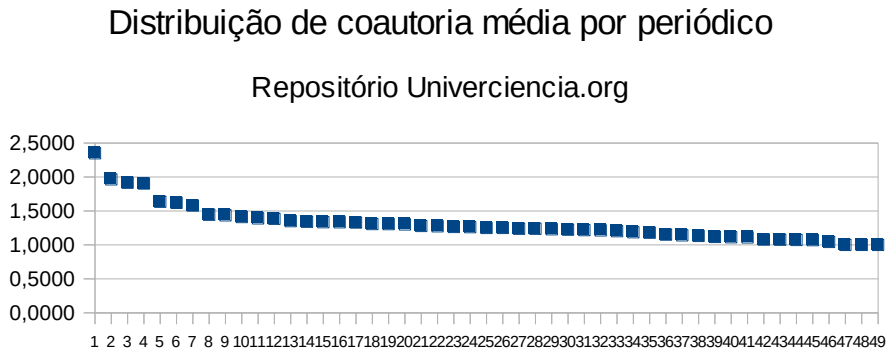

Figura 5: Distribuição da média de coautoria por periódico. Fonte: www.univerciencia.org

O resultado da figura 5 mostra um conjunto de 7 revistas acima da média de 1,5 autor por artigo, um grupo de 39 revistas na faixa acima de 1,0 e abaixo de 1,5 autor por artigo e 3 revistas com média de 1,0 autor por artigo. Vários fatores devem ser considerados quando da análise desses primeiros resultados, sobretudo considerando que temos revistas científicas que relacionam a área de Ciência da Comunicação com a área da Saúde, como é o caso da revista Interface, que possui a maior quantidade de artigos coletados pelo portal. Esse fato deve ser levado em consideração, pois o perfil de coautoria de áreas científicas varia conforme as diferentes estratégias que são utilizadas pelos pesquisadores quando da publicação de seus trabalhos. Sabemos que a área da Saúde possui média de coautores mais elevada que as áreas das Ciências Sociais Aplicadas, em geral. O que estamos de fato analisando aqui é a interface entre áreas e seus possíveis efeitos quando contextualizadas a partir do que podemos visualizar de seus

metadados.

A base de dados do portal Univerciencia.org tem se tornado, como podemos inferir a partir desses primeiros resultados que podemos apresentar da análise de seus metadados e da qualidade como têm sido trabalhados em suas fontes, um campo experimental fundamental quando se trata de avaliar padrões, comparar tendências e compreender melhor a própria dinâmica da comunidade científica que se reúne e circula pela área da Ciência da Comunicação.

Conclusão

O portal opera como um sistema federado de informações, no caso como um biblioteca digital federada específica com foco na área da Ciências da Comunicação. Logo, ele se torna um local extremamente privilegiado para estudos e análise relativos a essa área do conhecimento, pois todo o trabalho de sistematização e agregação da produção cientifica está ali concentrado, executado por dezenas, talvez centenas de pessoas, que se preocuparam em disponibilizar seus documentos em formato digital, criaram e produziram descritores de metadados no padrão Dublin Core, disponibilizaram em portais locais interoperáveis segundo os padrões OAI-PMH e permitiram que seus conteúdos fossem coletados livremente.

O Univerciencia pode ser visto como um centro agregador de redes da área da Ciências da Comunicação, que ali se entrecruzam e podem ser pesquisadas, visitadas, navegadas e analisadas a partir dos mesmos parâmetros. Ele amplia a mobilidade, a estabilidade e permutabilidade desses elementos naquilo que Bruno Latour (1998, p. 396) chamou de centros de cálculo:

> "... construir centros implica trazer para eles elementos distantes – permitir que os centros dominem a distância -, mas *sem* trazê-los "de verdade" - para evitar que os centros sejam inundados. Esse paradoxo é resolvido criando-se inscrições que conservem, simultaneamente, o mínimo e o máximo possível, através do aumento da mobilidade, da estabilidade ou da permutabilidade desses elementos. Esse meio-termo entre presença e ausência muitas vezes é chamado de *informação*. Quando se tem uma informação em mãos, tem-se a *forma* de alguma coisa sem ter a coisa em si."

Logo, o portal é um operador fundamental para viabilizar a amostra de dados mais abrangente possível da produção científica de sua área. A possibilidade de ampliar a expressão dos dados de referência para posteriores estudos, considerando a análise dos padrões de produção científica da comunidade da Ciência da Comunicação, é um fato que amplia o conhecimento bibliométrico e cientométrico que pode ser gerado a partir de seus resultados.

Também vale frisar o bom impacto que políticas públicas de fomento a periódicos científicos em formato aberto pode trazer para a construção de novas bases de dados específicas em outras áreas do conhecimento, permitindo que bancos de dados temáticos possam ser construídos e utilizados como referência de estudos e pesquisas, ampliando o conhecimento das dinâmicas, diferenças e semelhanças das comunidades científicas brasileiras. Os dados abertos em formatos padronizados facilitam a agregação e possibilidade de geração de novos estudos e avanços nas

pesquisas científicas relacionadas às áreas da Ciência da Informação e Bibliometria, em geral. Um passo fundamental para a constituição de material de base para o desenvolvimento dessas áreas no Brasil.

Planejamos estudos futuros que permitam a visualização e análise da estrutura e dinâmica das redes sociais de pesquisadores a partir de suas relações de coautoria nos artigos publicados, bem como avaliações de tópicos de interesse dos pesquisadores. Também pretendemos avançar no estudo da qualidade dos metadados coletados, dado que obtivemos bons resultados, mas apenas avaliamos neste estudo o campo que descreve os autores de um artigo, sendo importante avançarmos para compreender a dinâmica de uso e padronização dos outros campos disponíveis pelo padrão Dublin Core.

Artigo recebido em 05/07/2012. Aprovado em 03/09/2012.

Referências

BEAL, J. Metadata and data quality problems in the Digital Library. *Journal of Digital Information,* v. 6, n. 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. *Informação & Informação,* v. 15, n. esp., p. 110-129, 2010.

CHAVEZ, R. et al. Services make the repository. *Journal of Digital Information*, v. 8, n. 2, 2007.

COLE, T. W.; FOULONNEAU, M. *Using Open Archives Initiative Protocol for metadata harvesting.* [S.l.]: Libraries Unlimited, 2007.

DORNFEST, R.; BRICKLEY, D. Metadados. In: ORAM, A. (Org.). *Peer-to-peer:* o poder transformador das redes ponto a ponto. Editora Berkeley. 2001.

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. [S.l.: s,.n.], 2009.

_____; SOUTO, L. F. Dos sistemas de informação federados à federação de bibliotecas digitais. *Revista Brasileira de Biblioteconomia e Documentação,* v. 2, n. 1, p. 23-40, jan./jun. 2006.

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

LATOUR, B. *Ciência em ação:* como seguir cientistas e engenheiros sociedade afora. Campinas: Ed. Unesp, 1998.

_____. Razão que a razão desconhece: laboratórios, bibliotecas, coleções. In: PARENTE, A. (Org.). *Tramas da rede:* Sulina, 2004.

MARCONDES, C. H.; SAYÃO, L. F.; Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. *Ciência da Informação,* v. 30, n.3, set./dez. 2001.

NICHOLS, D. M.; McKAY, D.; TWIDALE, M. B. A lightweight metadata quality tool. *JCDL'08*, June 2008.

SAYÃO, L. F., Afinal, o que é biblioteca digital?. *Revista USP*, dez./fev. 2008-2009.

VAN DE SOMPEL, H.; LAGOZE, C. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, v. 6, n. 2, Feb. 2000.

WARD, J. Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems and Services*, v. 20, n. 1, p. 40-47, 2004.

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. São Paulo: Angellara, 2006. p. 82-114.

WITTEN, I. H.; BAINBRIDGE, D.; NICHOLS, D. M. *How to build a digital library*. 2nd ed. [S.l.]: Morgan Kauffmann Publishers, 2010.