Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 17, n. esp. 2 – III SBCC, p.136-158, 2012. ISSN 1518-2924. DOI:10.5007/1518-2924.2012v17nesp2p136

ANÁLISE DA DINÂMICA DE EVOLUÇÃO DAS REVISTAS CIENTÍFICAS E BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES EM ACESSO LIVRE NA ÁREA DAS CIÊNCIAS DA COMUNICAÇÃO: O CASO DO REPOSITÓRIO UNIVERCIENCIA.ORG

Dalton Lopes Martins[i]
Sueli Mara Soares Pinto Ferreira[ii]

Resumo: O crescimento do número de repositórios de revistas científicas e bibliotecas digitais de teses e dissertações em acesso aberto tem sido relatado em diversos estudos nos últimos anos. O objetivo deste artigo é avaliar, com enfoque específico na área da Ciência da Comunicação, como esse fenômeno tem acontecido e quais os seus impactos para essa comunidade científica. Apresentamos os principais resultados obtidos na construção de uma biblioteca digital federada, o repositório Univerciencia.org, o que nos permitiu avaliar como os repositórios em acesso aberto no padrão do protocolo OAI-PMH têm se difundido por entre as revistas e programas de pós-graduação da área, bem como padrões de uso e modos de operacionalização dos repositórios, fornecendo indícios que nos permitem avaliar como estes têm sido construídos.

Palavras-chave: Biblioteca digital federada. OAI-PMH. Metadados. Bibliometria.

EVOLUTION ANALYSIS OF OPEN ACCESS SCIENTIFIC JOURNALS AND THESES AND DISSERTATIONS DIGITAL LIBRARIES IN COMMUNICATION SCIENCE: THE UNIVERCIENCIA.ORG REPOSITORY CASE

Abstract: *The growth of the number of repositories of scientific journals and digital libraries of theses and dissertations in open access has been reported in several studies recently. The aim of this paper is to evaluate, with a specific focus in the field of Science Communication, how this phenomenon has happened and its impacts for this scientific community. We present the main results obtained in the construction of a federated digital library, the repository Univerciencia.org, which allowed us to assess how open access repositories in standard OAI-PMH have been widespread among the journals and postgraduate programs of the area, as well as usage patterns and modes of operation of repositories, providing evidence to enable us to assess how they have been built.*

Keywords: *Federated digital library. OAI-PMH. Metadata. Bibliometrics.*

[i] Universidade de São Paulo – daltonmartins@usp.br
[ii] Universidade de São Paulo – smferrei@usp.br

1 INTRODUÇÃO

A *Open Archives Initiative*[3] vem se estabelecendo como um modelo de transporte e compartilhamento de metadados desde a publicação do protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*)[4] em janeiro de 2001. Sendo um modelo de arquitetura da informação projetado para ampliar a interoperabilidade entre bibliotecas digitais e permitir a disseminação da informação de forma mais eficiente (COLE; FOULONNEAU, 2007, p. 3), tem sido utilizado como base no desenvolvimento de novos serviços de dados para essas bibliotecas.

A produção de novos serviços compreende 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. 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 do número de bibliotecas digitais que ofertam metadados de seu conteúdo seguindo os padrões do protocolo OAI-PMH (COLE; FOULONNEAU, 2007, p. 55) envolvendo 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; em 2012 já existem 914 periódicos em formato eletrônico registrados no Portal Scielo[5], e 95 bibliotecas de teses e dissertações no IBICT[6], além de contar com editais públicos para fomento de projetos de digitalização e disponibilização de acervos nesse formato (FERREIRA, 2009, p. 10).

É importante notar que a possibilidade de integração dos metadados disponibilizados nessas bibliotecas digitais permite aos pesquisadores estudarem grandes bancos de dados para diversas análises da produção científica. Dependendo da abrangência e da distribuição dessas

[3]Iniciativa dos Arquivos Abertos
[4]Protocolo para Coleta de Metadados da Iniciativa dos Arquivos Abertos
[5]Portal Scielo:http://www.scielo.br/
[6]IBICT – Instituto Brasileiro de Informação em Científica e Tecnológica

bibliotecas, podemos ainda considerar a hipótese de analisar toda ou pelo menos a maioria da produção científica de uma determinada área do conhecimento, considerando que suas principais instituições e pesquisadores publiquem sua produção nas bibliotecas digitais e revistas de acesso aberto.

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 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 JUNIOR; 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 para o seu estabelecimento como uma alternativa viável ao modelo de comunicação existente. As vantagens da interoperabilidade podem estimular o uso dos repositórios digitais nos blocos de construção de uma transformação no modelo de comunicação científica (LAGOZE; 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 ampla possibilidade de troca entre esses pesquisadores passa pela capacidade de interoperabilidade entre seus sistemas de informação escolhidos para a publicação do resultado de suas pesquisas.

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; 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 desse 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.

O objetivo deste artigo é analisar: como o surgimento de repositórios vem se desenvolvendo ao longo do tempo; a participação de pesquisadores; e a disponibilização de documentos em acesso aberto por meio do uso do protocolo OAI-PMH na área específica de Ciências da Comunicação. Para tanto, procedeu-se uma avaliação das revistas científicas e bibliotecas digitais de teses e dissertações coletadas pela biblioteca digital federada Univerciencia.org, especializada na área da Comunicação. O que buscamos aqui é mostrar, a partir de uma descrição da base de dados do repositório, como esse desenvolvimento vem ocorrendo, trazendo evidências que nos auxiliem a compreender como o acesso aberto por meio do protocolo OAI-PMH tem sido adotado pelas instituições e revistas científicas da área. Também nos interessa observar quais são os potenciais novos serviços de informação que poderiam ser gerados para a área dado seu grau de interoperabilidade alcançado.

2 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 Centro de Estudos em Design de Sistemas Virtuais da Escola de Comunicações e Artes da USP (CEDUS), que vem sendo implementado como parte dos projetos e estudos que tiveram início a partir de 2000. Segundo o próprio texto de apresentação do portal, disponível em seu site[7], ele é uma biblioteca digital federada que tem por objetivos:

a) constituir-se ponto de referência para difusão, busca, uso e compartilhamento da produção científica e acadêmica internacional em ciências da comunicação (cinema, rádio, televisão, relações públicas, publicidade, propaganda, jornalismo, editoração, cibercultura).

b) prover aos usuários uma interface única, convergente e integrada de acesso, busca e recuperação dos metadados da produção científica em ciências da comunicação (teses, dissertações, artigos de revistas, revistas, livros e capítulos de livros, trabalhos de eventos, anais, trabalhos de conclusão de curso, recursos educacionais abertos e outros) publicada em

[7]http://www.univerciencia.org/index.php/about

distintos países e fontes.

c) oferecer aos parceiros institucionais, gestores das fontes de informação indexadas, mecanismos para incrementar o acesso, a visibilidade, a organização e a geração de métricas e indicadores de gestão e governança de seus conteúdos.

d) contribuir com o desenvolvimento científico da área de Ciências da Comunicação, proporcionando espaço e dados para visualização e mapeamento do estado da arte da produção científica nacional e internacional.

Para cumprir seus objetivos, o Univerciencia.org busca coletar o conjunto mais abrangente possível de fontes de informação disponíveis, segundo seus critérios, sobre a área de Ciências da Comunicação. O portal considera como fonte de informação: portais, repositórios, revistas eletrônicas, bibliotecas digitais etc. de texto completo e em acesso aberto, divulgando distintas tipologias de recursos informacionais, cobrindo no todo ou em grande parte a área de Ciências da Comunicação, dentro daquilo que está disponível em acesso aberto.

O portal opera como um sistema federado de informações, no caso, como uma biblioteca digital federada específica com foco na área de Ciências da Comunicação. Logo, ele se torna um local extremamente privilegiado para estudos e análises relativas 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. Além disso, o portal torna-se, também, um local privilegiado para que possamos estudar como se desenvolve ao longo do tempo a estruturação e disponibilização de repositórios em acesso aberto, que abrangência tem para uma área do conhecimento, que desafios traz, que padrões revela e que experiências podemos analisar a partir do exercício de construção de tal portal.

O Univerciencia pode ser visto como um centro agregador de redes da área de 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, pg. 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 que está disponibilizada em acesso aberto. Vejamos agora de que amostra de dados estamos falando, apresentando a seguir as principais características do banco de dados do Univerciencia.

O portal possui 35.785 recursos informacionais coletados de 17 países diferentes, representando 98 fontes de informação de 68 instituições distintas. Vejamos como esses dados estão distribuídos na Tabela 1.

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 1. 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 2.

Tabela 2. Dados da distribuição dos recursos informacionais por tipos de recursos – Fonte. www.univerciencia.org – Acessado em 25/06/2011.

Tipo de Recurso	Itens	Fontes	Instituições	Países	%Itens
Artigos de revistas	16925	62	52	15	47,30%
Miscelânea	8214	5	5	4	22,95%
CBCC	3536	1	1	1	9,88%
Dissertação/Tese	3481	17	17	4	9,73%
Livro/Capítulo	2491	2	2	2	6,96%
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%**

Fica visível a abrangência da base de dados do portal a partir dos dados anteriormente apresentados, onde temos 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. No entanto, para o objetivo deste artigo iremos analisar apenas a dinâmica de evolução dos repositórios de artigos de revistas científicas e bibliotecas digitais de teses e dissertações, o que representa em torno de 58% dos recursos informacionais coletados pelo portal.

Vale lembrar que nossa escolha se deve, sobretudo, ao nosso interesse em avaliar como, ao longo do tempo, foram surgindo repositórios em acesso aberto passíveis de serem coletados em formato OAI-PMH na área da Ciência da Comunicação no Brasil, estando aqui representada pelas bibliotecas digitais de seus programas de pós-graduação e pelas revistas científicas nas quais os pesquisadores da área publicam o resultado de seu trabalho científico.

3 RESULTADOS

Entendemos que as escolhas de como realizar as etapas envolvidas na análise dos dados não são apenas escolhas técnicas, mas modos de tratamento da informação que irão influenciar em nossa capacidade de olhar e identificar padrões que possam nos fornecer pistas das tendências e fenômenos sociais daquilo que estudamos. Logo, vale frisar de antemão que o procedimento de análise envolveu diversas etapas de processamento dos dados de modo que obtivéssemos a maior abrangência e qualidade possível.

Para tanto, nos momentos em que foi possível, desenvolvemos a automatização de determinados processos produzindo programas de computador que auxiliassem sobretudo na identificação de padrões e campos utilizados dos metadados disponíveis. Outros processos foram realizados de modo manual, como a checagem do número de registros coletados pela biblioteca Univerciencia.org e os registros disponíveis nos sites originais de cada revista e biblioteca digital de tese e dissertação, quando existiam de modo independente. Essa composição de instrumentos de análise, ora automatizados, ora manuais, nos permitiu uma imersão nas escolhas e modos de tratamento dos dados de cada revista e biblioteca analisada, o que nos possibilita problematizar como esses modos podem ser vistos quando nosso objetivo é a integração dos dados e a construção de uma possível visão comum. Foi a partir

dessa visão que orientamos como essa seção foi construída e a maneira de apresentação de seus principais resultados.

Vejamos a seguir cada uma dessas etapas de coleta dos dados e seus principais resultados obtidos.

3.1 Revisão das fontes de informação

O processo de revisão das fontes de informação teve por objetivo checar se o número de registros coletados pelo Univerciencia.org estava de fato atualizado e de acordo com os últimos registros publicados pelas fontes. Nosso objetivo era garantir que não teríamos deixado de coletar material disponível por eventuais problemas técnicos ou checar dados não coletados mas que não estavam devidamente configurados para coleta pelas próprias fontes de informação. Além disso, encontramos alguns casos onde números antigos de revistas estavam sendo disponibilizados em formato aberto apenas nos últimos meses, logo, checamos as fontes também para garantir que havíamos coletado os últimos números disponibilizados.

Desse modo, para garantir que estávamos com os últimos dados disponíveis em formato aberto para coleta em cada fonte de informação, ao final de dezembro de 2011 rodamos o *harvester* para realizar a última coleta de metadados. Ao final dessa coleta contávamos com 142 fontes de informação em potencial, contabilizando um total de 54.347 registros no banco de dados.

No entanto, no meio desses registros havia diversos conteúdos que não seriam utilizados para o desenvolvimento de nosso trabalho de análise, dado que estávamos em busca apenas da produção científica de revistas e bibliotecas digitais de teses e dissertações brasileiras. Logo, filtrando o banco de dados com esse objetivo em vista, terminamos essa etapa com 49 fontes de informação de revistas científicas, contabilizando 9.864 registros; 12 bibliotecas digitais, contabilizando 1.961 registros. Portanto, trata-se de um universo de 61 fontes de informação e de 11.825 registros que utilizamos como base de nossas análises, conforme sintetizado na Tabela 3.

Tabela 3. Síntese da base de dados utilizada para análise.

Categorias	Fontes de Informação	Registros
Revistas científicas	49	9864
Bibliotecas de teses e dissertações	12	1961
Total	**61**	**11825**

Apresentamos nas Figuras 1 e 2, a seguir, a distribuição de frequência do número de registros coletados em nossa base de dados por fonte de informação, especificando o nome da fonte.

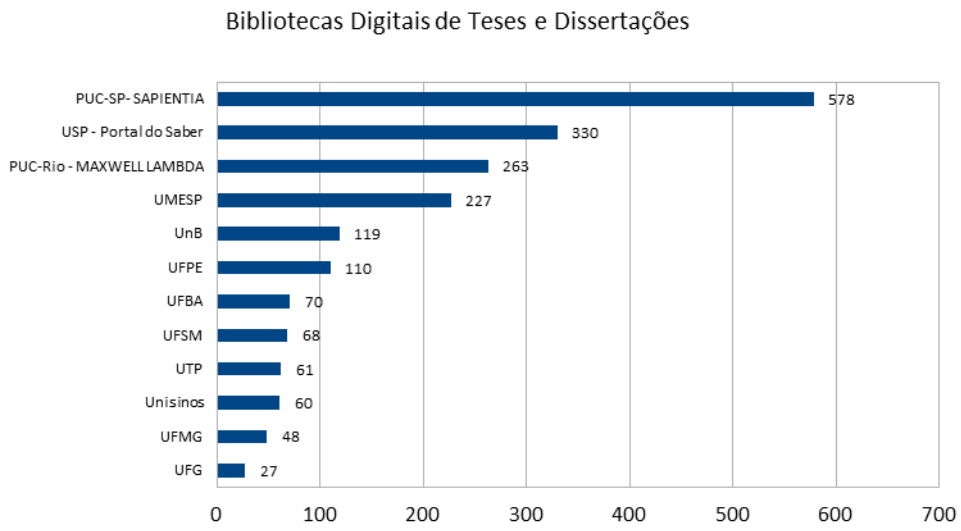

Figura 1. Distribuição da frequência do número de registros coletados por fonte de informação das bibliotecas digitais.

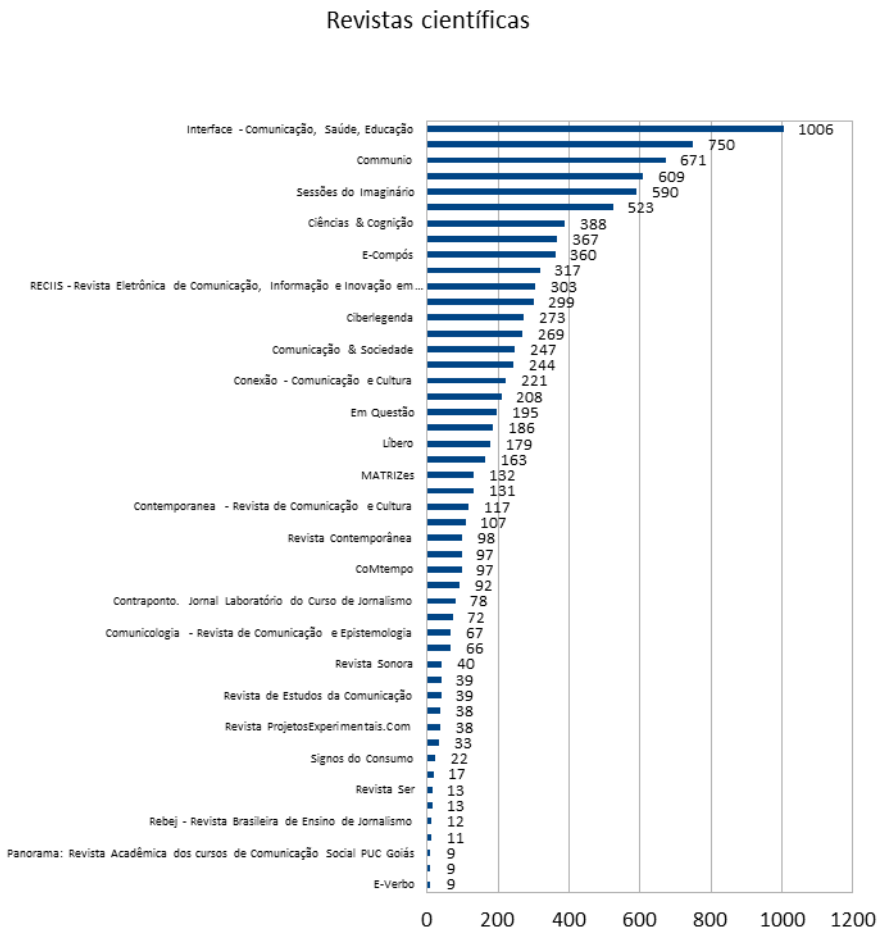

Figura 2. Distribuição da frequência do número de registros coletados por fonte de informação das revistas científicas.

3.2 Extração do banco de dados

Após termos identificado qual seria o universo de registros do banco de dados com os quais iríamos trabalhar, o próximo passo no tratamento dos dados foi extrair esses registros de modo que pudéssemos trabalhar apenas com os campos de metadados do formato Dublin Core, que eram de nosso interesse. Como nosso objetivo era o estudo da dinâmica de evolução dos repositórios em acesso aberto e sua potencial abrangência na área da Ciência da Comunicação, focamos na identificação, em cada registro, dos coautores de um artigo científico e dos participantes de uma banca de defesa de uma dissertação ou tese. Desse modo, estávamos apenas interessados nos campos que descrevem as pessoas envolvidas num determinado documento, além da data do campo, que consideramos ser a data de publicação do documento para podermos avaliar o processo dinâmico de evolução de nossos dados.

Todas as 61 fontes de informação com a qual estávamos trabalhando utilizaram apenas o formato de metadados Dublin Core simplificado. Logo, dos 15 campos disponíveis para

utilização pelo formato simplificado, apenas 3 eram de nosso potencial interesse para identificação das pessoas e datas, a saber: *creator, contributor e date.*

Os dados foram extraídos do banco de dados por fonte de informação e foram salvos em arquivos separadamente com o nome de cada fonte de informação. Após a separação dos dados, construímos um programa que fez automaticamente a extração desses 3 campos de cada registro, avaliando inicialmente se eles existiam e se estavam devidamente preenchidos. Foi por meio desse procedimento que pudemos avaliar quantos registros estariam de fato disponíveis para utilizarmos e quantos registros estariam com problemas técnicos de formatação ou com falta de informação suficiente para que pudéssemos considerá-los em nossas análises. Apresentamos na Tabela 4, a seguir, os resultados da validação dos metadados extraídos.

Tabela 4. Resultados do processo de validação dos metadados.

	Registros	
Categorias	**Coletados**	**Validados**
Revistas científicas	9864	9638 (97,7%)
Bibliotecas de teses e dissertações	1961	1953 (99,6%)
Total	**11825**	**11591 (98,0%)**

A extração e validação dos dados geraram uma redução mínima em relação ao número de registros coletados e aqueles que foram utilizados para nossas análises, conforme vemos na Tabela 4. Os principais problemas que encontramos no tratamento dos metadados foram de dois tipos: falta de informação no campo ou campos mal formados, ou seja, que não seguiam de modo estrito a forma como deveriam ser escritos segundo os parâmetros XML especificados pelo formato Dublin Core. Esses problemas foram detectados de modo esparso pelas diferentes fontes de informação e, como os dados ajudam a demonstrar, configuram-se mais como exceção no modo de constituição desses metadados do que regra.

A saída gerada por nosso programa de análise dos metadados criava um arquivo separado para cada fonte de informação contendo em cada linha os nomes das pessoas e a data extraída de cada registro. Esse arquivo permitia termos um controle do conjunto de nomes que apareceram em cada fonte de informação, bem como da data em que apareceram.

Além disso, esse formato possibilitava ao nosso programa contar de modo automático o número de autores, no caso das revistas científicas, que participaram de cada produção científica analisada. Já no caso das bibliotecas digitais, permitia contarmos o número de membros das bancas quando a biblioteca informava, ou registrávamos apenas a relação

orientador e orientando.

3.3 Normalização dos nomes

O processo de normalização dos nomes consistia em criarmos um identificador numérico único para cada pessoa que aparecesse em um determinado banco de dados. Inicialmente, compatibilizamos o modo com o qual cada nome era escrito, ou seja, garantindo que teríamos em cada registro uma estrutura que fosse composta pelo primeiro nome seguido dos possíveis sobrenomes de cada pessoa. Esse procedimento foi feito de modo automático, identificando e substituindo pelo modo correto de escrita os marcadores de separação de nomes em cada registro.

Após a unificação do método de escrita dos nomes, os mesmos foram organizados por ordem alfabética e comparados entre si por um programa de identificação automática. Os nomes iguais foram agrupados sob a mesma a identificação numérica. Para garantir maior consistência nos dados e afetar o menos possível a dinâmica de rede que investigávamos, foi realizada uma avaliação manual dos resultados de agrupamento dos nomes, permitindo identificarmos pequenas variações nos modos de escrita, melhorando assim a formação de grupos que antes poderiam ser considerados como nomes diferentes. Apresentamos na Tabela 5, a seguir, os resultados dessa etapa.

Tabela 5. Resultados do processo de normalização de nomes.

	Normalização de nomes	
Categorias	**Antes (nomes)**	**Depois (nomes)**
Revistas científicas	10393	9587 (92,2%)
Bibliotecas de teses e dissertações	2513	2465 (98,1%)

A normalização permitiu reduzirmos em praticamente 8% os nomes presentes na base de dados das revistas científicas e 2% na base de bibliotecas de teses e dissertações. A base de revistas apresenta uma quantidade maior de registros, logo, seria esperado que tivesse um universo maior de pessoas que de algum modo estivessem vinculadas à produção de um documento. Já na base de bibliotecas, o universo é menor não só devido ao número de documentos disponíveis, mas também pelo fato de que a entrada de uma pessoa nessa base exige que ela ao menos tenha defendido uma dissertação de mestrado, tese de doutorado ou participado de alguma banca, o que exclui potenciais alunos de graduação, pós-graduação *lato sensu*, entre outros, que podem estar envolvidos na produção de um artigo científico, por

exemplo.

Vale destacar que essa redução de nomes é fundamental para dar uma ideia da abrangência de pessoas de uma área do conhecimento que de algum modo se envolveram na produção científica registrada em acesso aberto em nosso banco de dados.

Apresentamos a seguir uma descrição das características dessa dinâmica, ou seja, como evoluíram ao longo do tempo os documentos e a própria formação dos repositórios de revistas científicas e de bibliotecas digitais de teses e dissertações coletados para análise no repositório Univerciencia.org.

3.4 Base de revistas científicas

A base de revistas científicas é constituída por 49 revistas, já apresentadas na Figura 2, e conta com 9.638 documentos válidos para análise. Vejamos na Figura 3, a seguir, como esses documentos são distribuídos ao longo do tempo.

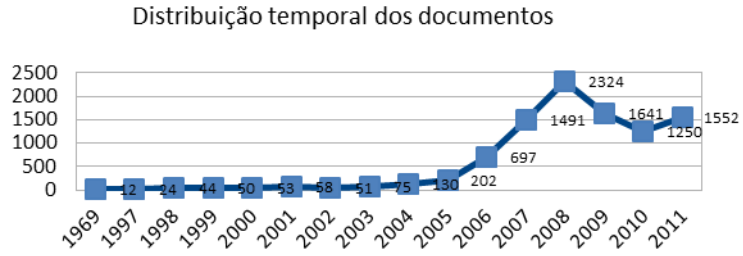

Figura 3. Distribuição temporal dos documentos na base de revistas científicas.

A distribuição tem início no ano de 1969, com um salto de datas até o ano de 1997, depois passa a ser regular em relação aos anos dos documentos, sendo o período regular de 15 anos de dados. Descartamos em nossa análise os dados relativos a 1969 para não prejudicar nossa escala temporal. É interessante notarmos o salto de crescimento na disponibilização de documentos datados sobretudo a partir de 2006, quando percebemos um aumento significativo na taxa de crescimento do número de documentos publicados, passando de 202, em 2005, para 697 em 2006, um crescimento de 345% em um ano, continuando a crescer de 2006 para 2007 e de 2007 para 2008, quando atinge seu pico. Podemos observar que esse crescimento parece acompanhar os dados relatados nos relatórios trienais da CAPES (2009), mostrando que do ano 2000 para 2009 houve um crescimento de 260% no número de

programas de pós-graduação no Brasil na área de Ciências da Comunicação.

No entanto, não podemos relacionar diretamente o crescimento do número de programas de pós-graduação com o crescimento da produção científica em formato aberto que pode ser coletada por uma biblioteca digital federada. Outro fator parece se somar ao crescimento do número de programas de pós-graduação na área: o maior incentivo à criação e ao desenvolvimento de revistas científicas em formato aberto no Brasil impulsionado por financiamento direto da CAPES e do CNPq. Juntos, esses órgãos passaram a investir em torno de R$ 4,8 milhões de reais a partir de 2007 (CAPES, 2006) em editais públicos visando incentivar a produção de revistas, inclusive mencionando explicitamente no texto dos editais ao longo desses anos que as revistas devem ser de acesso aberto, como vemos a seguir na declaração de objetivo do edital de 2008 (CNPq, 2008):

> Apoiar e incentivar a editoração e a publicação de periódicos científicos brasileiros, em todas as áreas do conhecimento, sendo considerado prioritário o apoio às revistas divulgadas por meio eletrônico, na Internet, em modo de acesso aberto, ou de forma impressa/eletrônica simultaneamente.

Observando na Figura 4, a seguir, a curva de revistas ativas em acesso aberto que estão publicando efetivamente ao longo dos anos de produção coletados pela biblioteca Univerciencia.org, percebemos claramente que há um salto de 17 revistas para 32 ativas de 2007 para 2008, representando um crescimento de 188% no número de revistas nesse período.

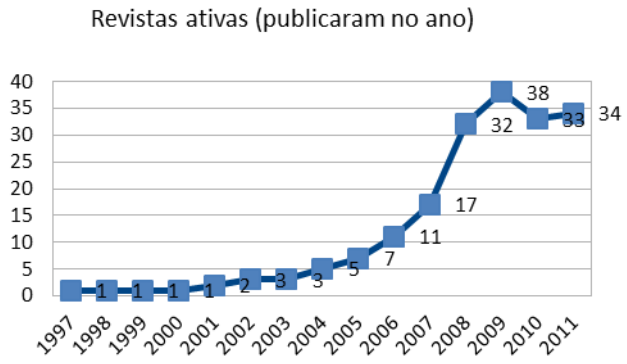

Figura 4. Distribuição temporal de revistas ativas na base Univerciencia.org.

Quando avaliamos em que ano essas revistas estão iniciando a publicação de artigos que foram coletados na Univerciencia, encontramos a curva apresentada na Figura 5, a seguir, que mostra como há um expressivo número de revistas iniciando suas atividades de 2007 para 2008, com um crescimento de 6 novas revistas em 2007 para 15 novas revistas em 2008, o que representa um crescimento de 250% no período.

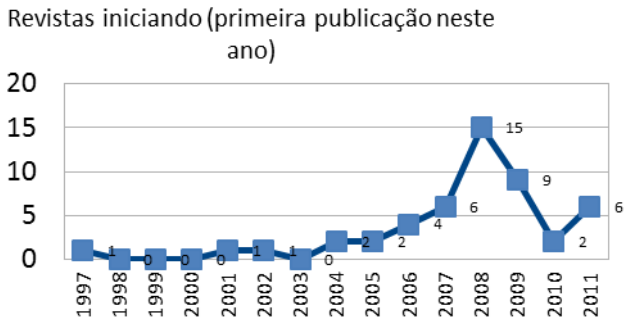

Figura 5. Distribuição temporal de novas revistas na base Univerciencia.org.

Quando observamos os resultados dos editais de 2006 (CNPQ, 2006) e 2007 (CNPQ, 2008), encontramos algumas importantes revistas da área da Ciência da Comunicação em termos de quantidades de documentos que foram coletados pela Univerciencia.org, como por exemplo, Interface – Comunicação, Saúde e Educação e Ciências & Cognição. Ao que tudo indica, a soma dos fatores envolvendo o crescimento no número de programas de pós-graduação na área das Ciências da Comunicação com o maior incentivo para o desenvolvimento de revistas científicas em acesso aberto, essencialmente ações de uma política pública científica, parece ter causado o impacto observado no crescimento de documentos e revistas científicas em acesso aberto que puderem ser coletadas pelo repositório Univerciencia.org.

Outro fator que vale mencionar é a queda no número documentos sendo publicados e no número de novas revistas surgindo na área entre os anos de 2008 e 2010, com uma recuperação nas taxas de crescimento percebida no ano de 2011. Não identificamos fatores explícitos, seja na dinâmica da política científica ou social, que pudessem explicar fatores que tenham levado a esse comportamento. É fato que observamos, naquilo que pode ser visto a partir das lentes proporcionadas por uma biblioteca federada, como a Univerciencia.org, uma redução no ritmo de publicação nas revistas que ali são indexadas. Outros fatores que extrapolam os limites deste trabalho podem ser levados em consideração como meio de explicar esse acontecimento, como o fato de pesquisadores estarem sendo levados a buscar outras revistas específicas de classificação Qualis que não são disponibilizadas em formato aberto e, portanto, que não foram coletadas na base Univerciencia, por exemplo.

O que é fundamental ter em questão quando da explicitação de nossas análises é a limitação com a qual estamos lidando, ou seja, estamos avaliando uma determinada biblioteca digital federada, o que de fato parece não refletir de maneira completa a dinâmica da área da

Ciência da Comunicação. Outros modos e locais de expressão da área são utilizados, porém, nosso objetivo aqui não é coletá-los e analisá-los, dado que estamos interessados exatamente na dinâmica dos repositórios de acesso aberto dessa comunidade científica.

Já a média ponderada geral da coautoria na base Univerciencia.org é de 1,43 autores por artigo, sendo que se considerarmos apenas o último ano analisado, em 2011, teremos uma coautoria média de 1,57 autores por artigo. Mostramos na Figura 6, a seguir, quais são as revistas que apresentam um padrão de coautoria maior que a média de 1,43, visando conhecermos quais são as revistas que mais contabilizam relações de colaboração entre pesquisadores por artigo publicado.

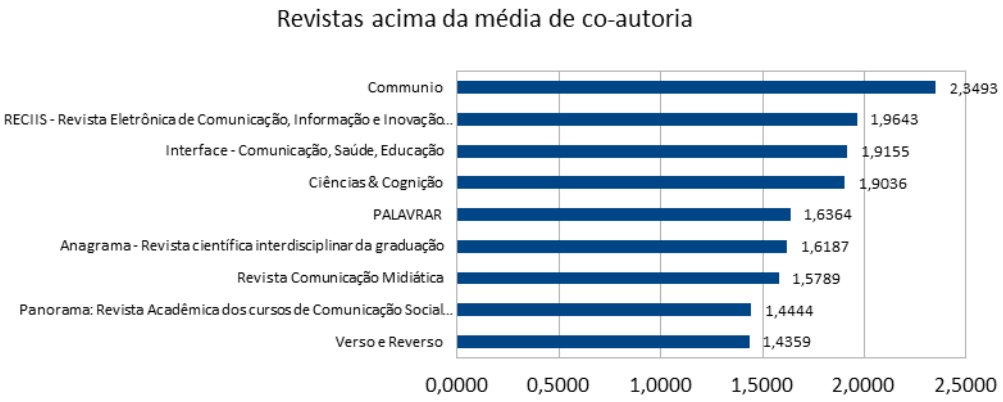

Figura 6. Distribuição de coautoria média por revistas.

É interessante observamos que das 4 principais revistas com maior média de coautoria por artigo, 3 delas de alguma maneira se relacionam com a área da saúde: RECIIS – Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Interface –, Comunicação, Saúde e Educação, e Ciências & Cognição, que é promovida e hospedada pelo Centro de Ciências da Saúde da Universidade Federal do Rio de Janeiro. Considerando que estamos falando de revistas multidisciplinares, que relacionam as questões da comunicação com as questões da saúde, percebemos aqui que encontramos estratégias de produção científica que parecem elevar a média de coautoria da área de Ciências da Comunicação.

Apresentamos na Tabela 6, a seguir, como se dá a distribuição total de documentos por faixa de coautoria na base Univerciencia.org.

Tabela 6. Distribuição de documentos por faixa de coautoria

Coautores	Documentos	%
1	7216	74,75%
2	1564	16,20%
3	439	4,55%
4	189	1,96%
5	112	1,16%
6	60	0,62%
7	34	0,35%
8	18	0,19%
9	9	0,09%
10	7	0,07%
11	1	0,01%
12	3	0,03%
15	1	0,01%
17	1	0,01%
Total	**9654**	**100,00%**

Nossos resultados indicam que em torno de 75% de toda a produção científica de revistas brasileiras na base Univerciencia.org é assinada por apenas um autor, e que 25% são distribuídos de 2 até o número máximo de 17 autores por artigo. De fato, os resultados afirmam que a maioria da produção científica da área disponibilizada em acesso aberto é assinada por apenas um único autor. No entanto, como podemos visualizar na figura 7, essa tendência parece estar mudando gradativamente, sobretudo do ano de 2008 para frente.

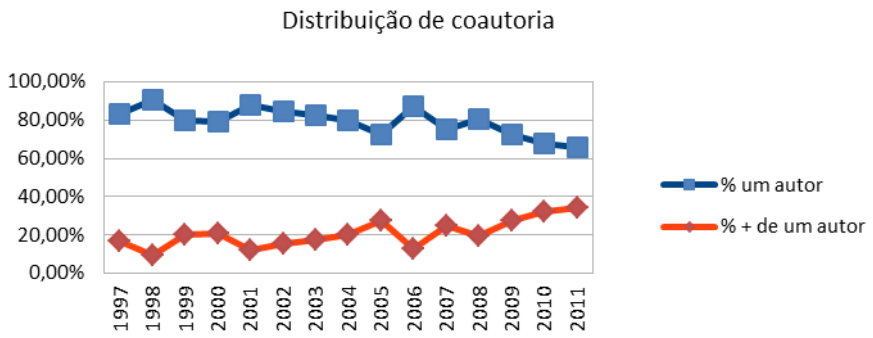

Figura 7. Distribuição de faixas coautoria por ano.

Analisando a figura 7, percebemos uma tendência de crescimento na coautoria de 2008 para 2011, levando a uma taxa de 20% dos artigos em coautoria para próximo de 40% em 2011. O perfil da área, a partir do que vemos no Univerciencia.org, parece estar se modificando.

Vejamos agora as características da base de bibliotecas digitais de teses e dissertações.

3.5 Base de bibliotecas digitais de teses e dissertações

A base de bibliotecas digitais de teses e dissertações é constituída por 12 bibliotecas, já apresentadas na figura 1 deste artigo, e conta com 1953 documentos válidos para análise. Vejamos na figura 8, a seguir, como esses documentos são distribuídos ao longo do tempo. A distribuição tem início no ano de 1986 e segue de modo regular até o ano de 2011, caracterizando um período amostral de 26 anos.

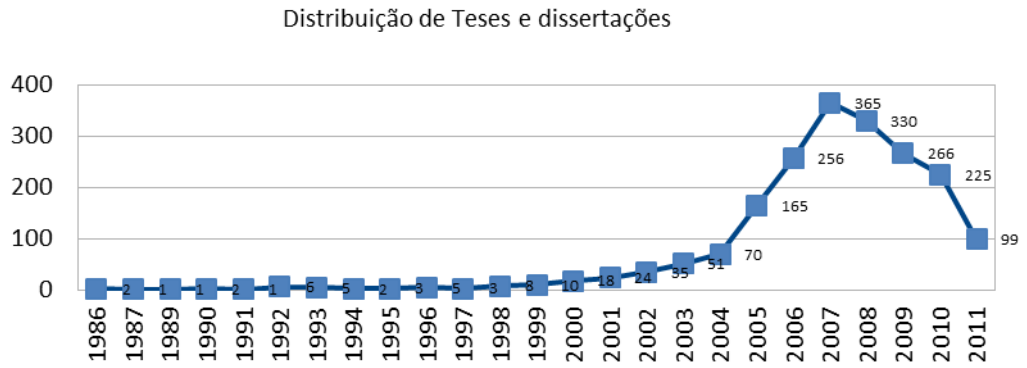

Figura 8. Distribuição teses e dissertações ao longo do tempo.

É interessante observarmos como a base entra num período de crescimento mais intenso a partir do início dos anos 2000, tendo um pico em sua taxa de crescimento do ano de 2004 para 2005, quando passa de 70 documentos para 165 documentos, representando um crescimento de 235% no período. Esse crescimento se mantém até atingir um pico em 2007, a partir de então caindo ano após ano até atingir 99 trabalhos no ano de 2011. É importante ressaltar que devemos considerar para o ano de 2011 que a extração final da base de dados foi feita no final de dezembro deste ano, sendo que possivelmente muitos documentos de teses e dissertações defendidas nos últimos meses do ano poderiam ainda não ter sido digitalizadas e disponibilizadas em suas bibliotecas digitais para coleta. No entanto, mesmo excluindo o ano de 2011 em nossas considerações, observamos uma queda de 38% na disponibilização de teses e dissertações nessas bibliotecas digitais para coleta do ano de 2007 para 2010.

Observando a listagem da CAPES (2009) para os programas de pós-graduação na área de Comunicação, encontramos um total de 41 programas disponibilizados por 40 instituições diferentes. Quando observamos nossas fontes de informação, como apresentado na figura 1, estamos tratando apenas de 12 bibliotecas digitais. Logo, estamos coletando em torno de 29% dos programas. Sem dúvida, vale mencionar que não coletamos mais dados devido aos

programas não disponibilizarem sua produção de teses e dissertações em formato aberto compatível com protocolo OAI-PMH, o que viabilizaria a construção de bibliotecas digitais federadas mais abrangentes e representativas da área.

Vejamos agora, na tabela 7, a abrangência temporal de cada uma das bibliotecas digitais coletadas.

Tabela 7. Abrangência temporal das bibliotecas digitais.

Nome	Ano inicial	Ano final	Dados de banca
Biblioteca Digital de Teses e Dissertação da PUC-SP- SAPIENTIA	1986	2011	NÃO
Biblioteca Digital de Teses e Dissertações da UnB	1992	2011	NÃO
Biblioteca Digital de Teses e Dissertações da USP - Portal do Saber	1994	2011	NÃO
Biblioteca Digital de Teses e Dissertações da UFBA	2001	2010	SIM
Biblioteca Digital de Teses e Dissertações da UFPE	2001	2009	NÃO
Biblioteca Digital de Teses e Dissertações da UTP	2002	2009	NÃO
Biblioteca Digital de Teses e Dissertações da UMESP	2002	2011	SIM
Biblioteca Digital de Teses e Dissertações da Unisinos	2003	2008	NÃO
Biblioteca Digital de Teses e Dissertações da UFSM	2004	2011	NÃO
Biblioteca Digital de Teses e Dissertações da PUC-Rio - MAXWELL LAMBDA	2005	2011	SIM
Biblioteca Digital de Teses e Dissertações da UFMG	2005	2011	SIM
Biblioteca Digital de Teses e Dissertações da UFG	2008	2011	NÃO

Observamos que do ano de 1986 até 1991, a biblioteca da PUC-SP é a única a ter em seu acervo documentos disponíveis em formato aberto para coleta, sendo seguida em 1992 pela UnB e em 1994 pela USP. Esse fator deve ser levado em consideração no número de documentos disponíveis em nossa base, sendo que a PUC sozinha é responsável por aproximadamente 30% dos documentos coletados. Até o ano de 2001 foram apenas essas três instituições que forneceram documentos que estão presentes em nossa base. A partir desse mesmo ano outras instituições entraram gradativamente disponibilizando seus documentos. Também vale notarmos que a Unisinos, a UFBA, a UFPE e a UTP não disponibilizaram documentos para coleta até o final do período, considerando o ano de 2011. Possíveis problemas em seus sistemas de informação ou questões de gestão podem ter inviabilizado esse trabalho. Vale considerar que isso produz efeitos na queda da produção científica que temos observado, dado que podemos estar observando não uma queda real de produção, mas sim uma não disponibilização desses documentos em acesso aberto para coleta.

Desse modo, o que observamos como uma queda na produção científica a partir do ano de 2007 deve ser contextualizado segundo essa perspectiva, ou seja, de que estamos olhando para apenas 29% dos programas de pós-graduação. Seria fundamental para uma possível conclusão mais abrangente observarmos o que tem acontecido na produção científica de outros programas de pós-graduação, inclusive pelo fato de que podemos estar diante de um processo de descentralização da produção científica na área das Ciências da Comunicação, onde o aumento no número de programas de pós-graduação tem permitido que vários

estudantes que antes se dirigiam a programas mais antigos e consolidados agora se direcionem para outras instituições e programas. Aqui, mais uma vez, apontamos para o fato de que estamos observando a produção científica que é disponibilizada em formato aberto para coleta e produção de bibliotecas digitais federadas que permitem o tipo de análise que ora apresentamos. Sem esse tipo de instrumento, torna-se praticamente proibitivo em termos logísticos construirmos uma visão mais ampla e integrada de uma área do conhecimento.

Outra informação importante e que pode ter grande impacto nos resultados de outros tipos de análises que podem ser feitas nesses repositórios, como por exemplo, análises de redes sociais entre participantes de uma banca de dissertação ou tese, é não termos uma uniformidade nos dados dos membros de bancas presentes em todas as bibliotecas digitais. Como apresentamos na Tabela 7, apenas 4 bibliotecas digitais mostram dados dos membros participantes das bancas de defesa, sendo elas a UFBA, UMESP, PUC-RIO e UFMG. As outras bibliotecas apenas apresentam dados dos orientadores e autores dos respectivos trabalhos. Isso, sem dúvida, reduz a abrangência do estudo de rede que poderíamos realizar se tivéssemos esses dados completos.

4 DISCUSSÃO E ENCAMINHAMENTOS FINAIS

A biblioteca digital federada se mostrou ao longo deste trabalho como um sistema de informação fundamental para a agregação de dados que, do contrário, se encontrariam dispersos em diversas fontes de informação e diferentes formatos, chegando mesmo a inviabilizar uma possível agregação de documentos na quantidade e intervalo de tempo que foram aqui analisados. Trabalhamos ao todo com mais de 11.800 documentos coletados e 61 diferentes fontes de informação. Na base de revistas científicas, obtivemos dados num intervalo de 15 anos, e na base de bibliotecas digitais, um intervalo de 26 anos, o que permitiu ao nosso estudo uma das maiores abrangências em termos de volume na área da Ciência da Comunicação no Brasil. Vale dizer que dificilmente teríamos conseguido agregar tal volume de dados neste intervalo de tempo se não tivéssemos um modo técnico homogêneo para que todas as fontes de informação pudessem se enquadrar, permitindo a disponibilização e coleta de seus dados. Sem dúvida, trata-se de um esforço coletivo da maior relevância para a área científica, permitindo que estudos como este presente possam ser realizados.

Além disso, vale aqui destacarmos os resultados em termos de qualidade dos

metadados que aqui foram coletados. Tivemos de descartar apenas 2% do número total de documentos que coletamos, dado que esses apresentavam problemas técnicos de formatação ou falta de informação para que pudessem ser considerados em nossa análise. Entendemos que essa perda de informação se tornou praticamente irrelevante para os resultados aqui apresentados. No entanto vale refletirmos aqui que quanto mais ricos e completos forem esses metadados, maiores possibilidades de análises poderiam ser feitas, tais como dados sobre a filiação de cada pesquisador, se são alunos, professores ou assistentes técnicos colaborando em uma pesquisa.

Em relação à abrangência de nosso estudo, vale ressaltar também a limitação que encontramos se estivéssemos interessados em caracterizar uma determinada área do conhecimento. Como demonstramos ao longo do trabalho, há hoje em torno de 41 programas de pós-graduação na área da Ciência da Comunicação e nós tínhamos à disposição apenas dados de 12 bibliotecas digitais de instituições que possuem programas de pós-graduação em Comunicação. Isso cria um efeito que é fundamental considerarmos como produto deste trabalho, sendo que o que conseguimos observar e relatar em nosso estudo é muito mais a dimensão alcançada por uma determinada federação de bibliotecas digitais do que propriamente a dinâmica social de uma determinada área do conhecimento. É importante evidenciarmos isso, pois não seria possível afirmarmos que nossos resultados servem para a área da Ciência da Comunicação como um todo, sabendo que muitos possíveis dados importantes não foram aqui coletados e analisados.

No entanto, vale aqui considerarmos alguns efeitos dessa análise. Se as revistas científicas, de modo geral, assim como as instituições que possuem programas de pós-graduação em uma determinada área do conhecimento disponibilizarem seu acervo de produção científica em formato aberto nos padrões do OAI-PMH, seguindo os mesmos padrões de qualidade e organização dos metadados que aqui relatamos, podemos afirmar que teríamos condições de construir uma federação de bibliotecas digitais que nos permitiriam coletar e analisar de modo abrangente a dinâmica social de formação da produção científica entre os pesquisadores dessa área. Desse modo, teríamos condições de generalizar o método de análise aqui apresentado com enfoque em uma determinada federação de bibliotecas digitais para uma área do conhecimento de modo geral. O que notamos a partir de nossos resultados é que o ambiente de uma biblioteca digital federada não apenas pode coletar e sistematizar a produção científica de uma área do conhecimento, mas pode influenciar de modo decisivo nos meios de organização dessa área, sendo uma fonte importante de indicadores, métricas e base para eventuais pesquisas que busquem entender as tendências

dessa área.

Já no que concerne a questões da dinâmica de seus sistemas de informação, notamos uma diferença importante nas bibliotecas digitais, dado que apenas uma minoria, em torno de 33% das bibliotecas, disponibilizam dados dos participantes das bancas de defesas. Seria fundamental que outras bibliotecas pudessem também aderir a esse formato, permitindo que tivéssemos uma análise mais completa da dinâmica de relação entre os pesquisadores que colaboram em torno desses eventos científicos. Em relação às revistas científicas, não percebemos nenhuma diferença significativa em termos do modo de organização de seus metadados.

Em relação a novos serviços de informação, há um enorme potencial a ser explorado em termos de possibilidades de análise dos dados gerados por uma federação de bibliotecas digitais. Nosso trabalho aponta diretamente para estudos e mapeamentos de análise das redes sociais de relação entre os pesquisadores, seja na coautoria de artigos para revistas científicas, seja na participação em bancas de defesas de teses e dissertações. Além disso, outros campos dos metadados podem ser utilizados para analisar as principais palavras-chave, os principais temas baseados nos títulos dos trabalhos, o modo de relação institucional baseado na filiação dos pesquisadores, a mobilidade de pesquisadores por entre instituições considerando a série histórica dos dados, entre outros estudos que poderiam ser aqui listados.

De modo geral, a biblioteca digital federada parece exercer de forma privilegiada esse papel de centro de cálculo, permitindo a mediação entre centros distribuídos que divulgam sua produção científica em acesso aberto seguindo o protocolo OAIP-PMH, além de darem suporte a uma integração singular desses dados, viabilizando análises bibliométricas, cientométricas e mesmo webométricas.

REFERÊNCIAS

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.

CAPES. **Relatório de avaliação trienal da área da Ciências Sociais Aplicadas I 2007-2009**. Disponível em: http://tinyurl.com/6p92p96. Acesso em: 08 mar. 2012.

CAPES. **Capes se associa ao CNPq no financiamento de revistas científicas nacionais**, 2006. Disponível em: http://tinyurl.com/87szckl. Acesso em: 08 mar. 2012.

CAPES. **Relação de Cursos Recomendados e Reconhecidos 2007-2009**. Disponível em: http://tinyurl.com/74mae6l. Acesso em: 08 mar. 2012.

CNPq. **Edital MCT/CNPq/MEC/CAPES nº 58/2008**. Disponível em:http://www.cnpq.br/resultados/2008/058.htm. Acesso em: 08 mar. 2012.

CNPq. **Edital MCT/CNPq/MEC/CAPES No 16/2007**. Disponível em: http://www.cnpq.br/resultados/2007/docs/016.pdf. Acesso em: 08 mar. 2012.

COLE, T. W.; FOULONNEAU, M. Using Open Archives Initiative Protocol for metadata harvesting. Libraries Unlimited. 2007. 208p.

FERREIRA, S. M. S. P.; MORAIS, M. H.; MUCHERONI, M.; PEREZ, J. Estudo sobre como autores de artigos de revistas de Ciências da Comunicação verbalizam seus objetos de estudos em termos de palavras chave. **Em Questão**, Porto Alegre, v. 15, n. 2, p. 151-167. 2009.

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

LATOUR, Bruno. **Ciência em ação**: como seguir cientistas e engenheiros sociedade afora. Editora Unesp. 1998. 438p.

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

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.