ELISABETE FERREIRA

# UM MÉTODO DE COLETA E CLASSIFICAÇÃO DE METADADOS DE PRODUÇÃO CIENTÍFICA EM REPOSITÓRIOS DIGITAIS INSTITUCIONAIS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas da Universidade Federal do Paraná
Orientador: Prof. Dr. Marcos Sfair Sunye

CURITIBA

2016

Catalogação na Publicação
Sistema de Bibliotecas UFPR
Karolayne Costa Rodrigues de Lima - CRB 9/1638

Ferreira, Elisabete
Um método de coleta e classificação de metadados de produção científica em repositórios digitais institucionais / Elisabete Ferreira – Curitiba, 2016.
63f.: il. color.

Orientador: Prof. Dr. Marcos Sfair Sunyé
Dissertação (Mestrado em Informática) – Setor de Ciências Exatas, Programa de Pós-graduação em Informática, Universidade Federal do Paraná.

1. Acesso aberto 2. Metadados - Colheita automatizada - Repositórios digitais 4. Mineração de dados (Computação) 5. Publicações científicas I.Título.

CDD 025.0637

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS EXATAS
Programa de Pós Graduação em INFORMÁTICA
Código CAPES: 40001016034P5

## TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **ELISABETE FERREIRA**, intitulada: "**UM MÉTODO DE COLETA E CLASSIFICAÇÃO DE METADADOS DE PRODUÇÃO CIENTÍFICA EM REPOSITÓRIOS DIGITAIS INSTITUCIONAIS**", após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 05 de Julho de 2016.

Prof MARCOS SFAIR SUNYE
Presidente da Banca Examinadora (UFPR)

Prof MARCOS DIDONET DEL FABRO
Avaliador Interno (UFPR)

Prof SONIA ELISA CAREGNATO
Avaliador Externo (UFRGS)

MINISTÉRIO DA EDUCAÇÃO - UNIVERSIDADE FEDERAL DO PARANÁ - Setor de Ciências Exatas - Coordenação de Pós-Graduação em Informática

*Este trabalho é dedicado àqueles que, mesmo nas adversidades, não deixam de sonhar.*

## AGRADECIMENTOS

Agradeço, primeiramente, à Deus, Jesus e Nossa Senhora pela proteção e energia que me permitiram concluir este trabalho.

Agradeço aos meus familiares, especialmente à minha mãe, Nilde, pelo incentivo dispendido em todo o período em que estive na universidade e ao meu padrasto Paulo José (*in memoriam*), que onde quer que esteja, nunca deixou de me amar e confiar em mim.

Aos meus colegas, Carlos André, Clariane, Edgar, Igor e Renato pela força e paciência em nossos grupos de estudo e, também nos momentos de lazer.

Aos meus colegas do Departamento de Informática da UFPR: Andrea, Jucélia, Letícia, Rafael e Raquel, pelo auxílio e amizade.

Aos bolsistas Aluísio e Bruno Zanette, pela ativa participação no desenvolvimento deste trabalho.

Aos monitores Guilherme e Ricardo, imprescindíveis na tarefa de trazer à tona conceitos anteriormente já vistos.

Ao meu "mestre" Marcos Sunye, pelo esforço e dedicação nesta importante caminhada.

Aos meus colegas do Sistemas de Bibliotecas da UFPR: Fabiane, Karol, Ligia, Lucas, Paula e Suzana, pelo apoio e participação.

Enfim, agradeço a todas as pessoas que contribuíram, direta ou indiretamente nesta etapa decisiva em minha vida.

*"Jesus, porém, respondendo, disse-lhes: Em verdade vos digo que, se tiverdes fé e não duvidardes, não só fareis o que foi feito à figueira, mas até se a este monte disserdes: Ergue-te, e precipita-te no mar, assim será feito; E, tudo o que pedirdes em oração, crendo, o recebereis."*
*(Bíblia Sagrada, Mateus, 21, 21–22)*

## RESUMO

A agregação da produção científica em um único ambiente digital institucional permite às instituições gerar indicadores internos de produção científica e tecnológica, realizar estudos através da aplicação de ferramentas de mineração de dados, bem como apoiar a implementação de políticas de gestão. Embora as instituições acadêmicas sejam as grandes produtoras de conhecimento científico, enfrentam dificuldades para identificar, agregar e quantificar o próprio conhecimento produzido em seus ambientes digitais e, por conseguinte definirem critérios precisos para planejamento e distribuição de recursos que fomentem a produção científica por parte de seu corpo docente. Este trabalho apresenta uma metodologia para carga automática de metadados e artigos científicos disponibilizados em acesso aberto e dispersos em periódicos científicos, em Repositórios Digitais Institucionais obtidos por meio de extração de dados contidos nos currículos dos docentes da Universidade Federal do Paraná, registrados na Plataforma Lattes, além de auxiliar a instituição no planejamento dos custos necessários para manutenção de seu ambiente digital, através da obtenção do volume de produção científica a ser armazenado em seu repositório digital institucional. Para efeito da implementação da proposta foi desenvolvido um conjunto de componentes para mineração de artigos científicos produzidos e disponibilizados em acesso aberto na plataforma DSpace.



**Palavras-chave**: Acesso aberto. Colheita automatizada. Metadados. Mineração de dados. Publicações científicas. Repositórios digitais.

## ABSTRACT

The aggregation of scientific production in a single institutional digital environment allows institutions to generate internal indicators of scientific and technological production, conduct studies through the application of data mining tools as well as support the implementation of management policies. Although the academic instituitons have been the great cientific knowledge generators, they face dificulties in identifying, aggregating and quantifying their knowledge in their digital environments, and as a consequence to define precise criteria for planning and distributing resources that to encourage the cientific production by their researchers. This work proposes a methodology for the automatic loading of metadata and open access scientific articles, spread out in scientific journals in Institutional Digital Repositories, obtained through extraction of data contained in the curricula registered in the Lattes Platform of teachers of Federal University of Paraná. A further objective is to assist the institution for planning the costs required to maintain their digital environment by obtaining the volume of scientific production to be stored in its institutional digital repository. For the purpose of implementation, a set of components was developed for data mining of scientific articles produced and made available in open access on DSpace.



**Keywords**: Automated harvesting. Data mining. Digital repositories. Metadata. Open access. Scientific publications.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE SIGLAS

| | |
|---|---|
| API | – Application Programming Interface |
| CAPES | – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| CNPq | – Conselho Nacional de Desenvolvimento Cientifico e Tecnológico |
| CNRI | – Corporação para Iniciativas de Pesquisa Nacional |
| DCMI | – Iniciativa de Metadados Dublin Core |
| DiVA | – Digitala Vetenskapliga Arkivet |
| DOAJ | – Directory of Open Access Journals |
| DOI | – Digital Object Identifier |
| IDF | – *International DOI Foundation* |
| ISSN | – International Standard Serial Number |
| OAI-PMH | – Open Archives Initiative Protocol for Metadata Harvesting |
| ORCID | – Open Researcher and Contributor ID |
| PDF | – Portable Document Format |
| RDI/UFPR | – Repositório Digital Institucional da Universidade Federal do Paraná |
| RDI | – Repositório Digital Institucional |
| UFPR | – Universidade Federal do Paraná |
| XML | – Extensible Markup Language |

**SUMÁRIO**

# 1 INTRODUÇÃO

As instituições de ensino superior são as responsáveis por produzir material ou conhecimento científico em forma de artigos, relatórios, congressos, eventos de extensão e outros.

Segundo (SANTOS et al., 2005), há um conjunto expressivo de indicadores empregados para a análise da produção científica, que podem ser agrupados em:

- **Indicadores de produção científica**: construídos a partir da contagem do número de publicações por tipo de documento (livros, artigos, publicações científicas, relatórios etc.). Os indicadores de produção científica podem ser quantificados por instituição, área de conhecimento, país, e outros;

- **Indicadores de citação**: quantificam o número de citações recebidas por uma publicação de artigo em periódicos científicos. Citações são a forma mais usual de medir o impacto de uma publicação bem como, de atribuir crédito ao autor;

- **Indicadores de ligação**: quantificam as as ocorrências de co-autoria em uma publicação científica, o número de citações e palavras. São aplicados para a elaboração de mapas de estruturas de conhecimento e de redes de relacionamento entre pesquisadores, instituições e países. São utilizadas técnicas de análise estatística de agrupamentos para sua construção.

No universo de publicação de trabalhos científicos, acesso aberto é o conceito atrelado a publicações na internet que permitem a leitura, cópia, distribuição ou reutilização para propósitos lícitos, sem barreiras técnicas, financeiras ou legais, garantidos porém os direitos morais e patrimoniais de autoria (*BUDAPEST OPEN ACCESS INITIATIVE*, 2002).

A filosofia de acesso aberto refere-se ao movimento observado nos últimos anos em direção ao uso de ferramentas, estratégias e metodologias que denotam uma nova maneira de demonstrar um novo processo de comunicação científica.

Embora as instituições acadêmicas sejam as grandes produtoras de conhecimento científico, estas mesmas instituições têm dificuldades para agregar e quantificar o próprio conhecimento produzido e, por conseguinte definirem critérios precisos para planejamento e distribuição de recursos que fomentem a produção científica por parte de seu corpo docente.

Outra dificuldade das instituições é o monitoramento de sua produtividade intelectual por meio de indicadores, bem como planejamento adequado do processo de

arquivamento e preservação de materiais digitais a longo prazo, pela falta de uma ferramenta que precise efetivamente os custos necessários para implantação e manutenção de seus ambientes digitais.

Com a informatização das instituições acadêmicas, surgiram as Bibliotecas Digitais, cuja relevância nos dias atuais é incontestável. São elas as responsáveis por agregar, selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir e preservar trabalhos digitais de uma instituição de modo a estarem sempre prontos e economicamente disponíveis para a comunidade (LANGIANO, 2005). Além disso, por ampliar o acesso aos resultados de pesquisa de uma instituição, as Bibliotecas Digitais beneficiam profissionais e estudantes que se utilizam dos seus recursos em tarefas de ensino e aprendizagem.

Para disseminar o conhecimento produzido na forma de artigos científicos de uma instituição, os resultados são publicados em periódicos científicos, pois estes são considerados como o modo mais rápido e economicamente viável de difundir o conhecimento científico, os resultados de trabalhos de pesquisa e o que estes trabalhos representam para a comunidade (BROFMAN, 2012).

Existem diferentes formas de submissão e publicação de textos cientificos e, de acordo com Brofman (2012), publicações científicas, "objetivam divulgar a pesquisa científica para a comunidade, de forma que permita que outros possam utilizá-la e avaliá-la sob outras visões". Os autores de produções científicas são orientados a disponibilizarem seus trabalhos em um formato eletrônico normalizado em pelo menos um repositório de acesso aberto conforme a *Berlin Declaration on Open Access* (2003).

As publicações científicas submetidas a um periódico tradicional passam pelos passos *preprint*, que significa que a mesma foi submetida e está no processo de revisão por pares (*peer review*), e *post-print*, que significa que a mesma foi aceita para publicação e atualizada pelo autor com as sugestões obtidas na revisão por pares.

Existe ainda o termo *e-print*, surgido a partir da fundação do repositório ArXiv em 1991 por Paul Ginsparg (LUCE, 2001), que é a submissão pelo próprio autor em meios eletrônicos, observadas as políticas de acesso que acompanham a publicação. *E-print* inclui qualquer trabalho eletrônico distribuído pelo autor fora do ambiente tradicional de publicação (APS NEWS, 1998), visando, por exemplo, o compartilhamento prévio de seus resultados com colegas.

Ocorre, porém, que as Bibliotecas Digitais apresentam deficiência na coleta, seleção e agregação de itens de produção científica publicados nestes periódicos. Assim sendo, muitas delas focam unicamente na disponibilização de parte da produção científica de seus programas de ensino, na forma de monografias, teses e dissertações.

Diante desse cenário e no âmbito da consolidação de interesses em torno desse tema, expõe-se o problema de que como os artigos científicos de instituições estão dispersos

em periódicos científicos, o acesso e identificação a eles é muitas vezes dificultado à comunidade e até à própria instituição produtora do trabalho científico, pois é necessário realizar consultas aos diversos periódicos existentes para saber com precisão onde se encontra a produção científica específica desejada. E, ainda, as instituições também não dispõem de informações sobre quanto do seu corpo docente está sensibilizado quanto à disponibilização de sua produção científica em acesso aberto.

Para solucionar este problema e para a representação fiel da produção científica em Bibliotecas Digitais, a presente dissertação apresenta uma proposta de metodologia para carga automática de artigos científicos em Repositórios Digitais Institucionais obtida por meio de extração de dados contidos nos currículos dos docentes de uma instituição, registrados na base brasileira de currículos de pesquisadores e grupos de pesquisa Plataforma Lattes (2012).

A publicação em repositórios digitais promove a transferência de conhecimento que, segundo Barreto (2012), permite, com o emprego de uma tecnologia adequada, construir ou modificar um produto ou serviço, seu processo de transformação ou comercialização. Ainda segundo o mesmo autor, à toda tecnologia é associada uma considerável quantidade de informação que, quando assimilada pelo indivíduo, grupo ou sociedade gera um conhecimento que permite ou não adotar determinada técnica.

Exemplos de agregação de produções científicas em acesso aberto são o repositório ArXiv da Cornell University Library que oferece a autores a possibilidade de armazenar e compartilhar sua produção científica em um servidor eletrônico altamente automatizado para artigos de pesquisa (*ARXIV*, 2016) e o repositório multidisciplinar francês Hal criado para o arquivamento e disseminação de publicações científicas em acesso aberto de instituições francesas ou estrangeiras, públicas e laboratórios privados (*HAL*, 2016).

Para efeito da implementação prática (capítulo 8) foi realizado um estudo dividido em 4 partes: (1) obtenção e tratamento de metadados provenientes da Plataforma Lattes; (2) desenvolvimento de um *script* para mineração de artigos científicos produzidos e disponibilizados em acesso aberto; (3) definição de um software para carga e conversão dos metadados em formato *Dublin Core* e (4) alimentação do repositório digital institucional através da importação dos metadados obtidos. Para efeito do estudo foi escolhido o Repositório Digital Institucional da Universidade Federal do Paraná (RDI/UFPR), que utiliza a versão 5.4 da plataforma DSpace.

## 1.1 OBJETIVOS

Este trabalho tem como objetivo geral definir e implementar um método de coleta e classificação de metadados e artigos científicos para carga automática no RDI/UFPR, dos artigos produzidos pelos docentes da instituição, que foram publicados em acesso

aberto.

No mesmo sentido, os objetivos específicos são:

- Desenvolver uma ferramenta para identificar artigos científicos produzidos e publicados em periódicos de acesso aberto;
- Definir critérios para importação e tratamento de metadados de artigos científicos, relacionando-os com os dados funcionais dos docentes da UFPR listados na Plataforma Lattes;
- Desenvolver critérios para capturar de forma automática os PDFs constantes em periódicos de acesso aberto;
- Converter e, migrar metadados coletados em formato *Dublin Core*, para o RDI/UFPR;

## 1.2 ORGANIZAÇÃO DESTE ESTUDO

O Capítulo 2 trará o conceito de "Bibliotecas e Repositórios Digitais institucionais", sua importância, origem e características.

O Capítulo 3 discorrerá sobre metadados, finalidades e características, além de sua importância na recuperação de dados e interoperabilidade entre repositórios digitais.

No Capítulo 4 são conceituados identificadores persistentes, DOIs e *handles*, seus usos e finalidades para a preservação de materiais digitais a longo prazo.

O Capítulo 5 apresentará métodos comumente utilizados para atualização de repositórios digitais a partir de sistemas capazes de promover buscas e captura de metadados através de protocolos de interoperabilidade participantes da Iniciativa de Arquivos Abertos (OAI) e as dificuldades quanto à identificação de publicações científicas produzidas e disponibilizadas de acordo com o movimento de acesso aberto.

O Capítulo 6 apresentará diversas bases de metadados relacionadas a publicações científicas, tais como a Plataforma Lattes, CrossRef e SHERPA/RoMEO, e os principais métodos para obtenção de informações a partir destas bases.

O Capítulo 7 discutirá tópicos sobre direitos autorais e licenciamento de objetos digitais. Serão apresentadas as principais formas de licenciamento e suas principais características, entre elas as que sinalizam como uma obra digital pode ser utilizada e/ou compartilhada.

Finalmente, no Capítulo 8, este trabalho propõe uma nova metodologia para coletar publicações científicas e trazê-las para os repositórios digitais institucionais de acordo com a Iniciativa de Acesso Aberto. Será apresentado o desenvolvimento prático do trabalho cujas tarefas realizadas foram: carga e análise de arquivos Lattes no formato

Extensible Markup Language (XML), cruzamento de dados da Plataforma Lattes e dados funcionais dos docentes da instituição, identificação de artigos em acesso aberto, extração de metadados, processamento e carga desses metadados no RDI/UFPR.

O Capítulo 9 conclui o trabalho, discutindo os objetivos propostos, os resultados alcançados e os meios utilizados para alcançá-los e, ainda, alguns tópicos para trabalhos futuros que possam dar continuidade a este trabalho.

# 2 REPOSITÓRIOS DIGITAIS

Repositórios digitais são coleções de materiais em meio digital, organizadas e disponibilizadas para acesso *online*.

Os materiais ou objetos digitais disponibilizados em uma Biblioteca Digital podem derivar-se de cópias digitais de materiais existentes em meio físico como livros, impressos, manuscritos e outros, e/ou referir-se a objetos só existentes em meio digital como fotografias digitais, e-books, etc.

Para obter um processo confiável de identificação e localização de um objeto digital, além da utilização de uma rigorosa metodologia científica para a geração do conhecimento, é importante que os resultados obtidos pelas pesquisas acadêmico-científicas de uma instituição sejam divulgados em repositórios digitais de acesso aberto atrelados a um identificador persistente, que nada mais é do que um nome exclusivo para um objeto digital, e que independe de sua localização ou formato, garantindo assim que o mesmo seja acessível independente de mudanças físicas e tecnológicas (SAYÃO, 2007a).

Repositórios digitais propagaram-se em decorrência do aumento do custo de publicação em periódicos científicos em meados de 1975 a 1995, com o ressentimento de cientistas de países periféricos com o modelo oligopolizado das editoras científicas. Sendo assim, a partir de 1990 inicia-se um movimento, denominado "acesso aberto", para disponibilização do saber produzido pela ciência, possibilitado pelo avanço em larga escala da Internet (LIMA, 2009).

O conjunto de publicações, denominado "literatura científica", permite expor o trabalho dos pesquisadores de uma instituição ao julgamento constante de seus pares, em busca do consenso que lhes confere confiabilidade (*HAL*, 2016).

Nos Repositórios Digitais Institucionais (RDIs) é possível o armazenamento e difusão desta literatura científica, que engloba artigos de periódicos, teses e dissertações, e livros.

Para Setenareski (2013), "Com a oferta de *software* livre com código fonte aberto para a criação de repositórios e a difusão do movimento de acesso aberto, muitas centenas de instituições de ensino e pesquisa no mundo foram criando repositórios institucionais para garantir a difusão dos conteúdos científicos gerados dentro delas."

A adoção de um RDI por uma instituição significa mais do que apenas registrar e disseminar sua produção científica, mas sim envolve também todos os mecanismos de gestão e maximização da visibilidade da produção científica da instituição, além de que objetos digitais armazenados em um repositório estão, tecnicamente garantidos, no que se refere à preservação digital (CASTRO et al., 2009).

Repositórios digitais como um ambiente recente podem sofrer mudanças conceitu-

ais ao longo do tempo, porém, devem, sempre, manter-se fiéis ao princípio da preservação da memória a longo prazo. Alguns elementos essenciais devem ser abordados em ambientes científicos digitais, principalmente em repositórios digitais: Ferramentas de busca, Metadados, Interoperabilidade, Políticas, Preservação, Acessibilidade e Usabilidade. (CAMARGO et al., 2009).

Os RDIs são intensamente utilizados para fomentar a produção científica proveniente de atividades de pesquisa e ensino, bem como oferecem suporte às mesmas, pois conforme Leite (2009):

- melhoram a comunicação científica interna e externa à instituição;
- maximizam a acessibilidade, o uso, a visibilidade e o impacto da produção científica da instituição;
- retroalimentam a atividade de pesquisa científica e apoiam os processos de ensino e aprendizagem;
- apoiam as publicações científicas eletrônicas da instituição;
- contribuem para a preservação dos conteúdos digitais científicos ou acadêmicos produzidos pela instituição ou seus membros;
- contribuem para o aumento do prestígio da instituição e do pesquisador;
- oferecem insumo para a avaliação e monitoramento da produção científica;
- reúnem, armazenam, organizam, recuperam e disseminam a produção científica da instituição.

## 2.1 PROPRIEDADES DE REPOSITÓRIOS INSTITUCIONAIS

Repositórios digitais institucionais de acesso aberto, compreendem, necessariamente a natureza acadêmico-científica das publicações científicas nele armazenadas, e são distintos devido as seguintes propriedades (LEITE, 2009):

- institucionalmente definidos: restringem-se apenas ao conteúdo relacionado de uma instituição;
- científicos ou academicamente orientados: apresentam a produção intelectual da instituição;
- cumulativos e perpétuos (permanentes): garantem a disponibilidade ao seu conteúdo;
- abertos e interoperáveis: permitem a comunicação com outros repositórios;

- não efêmeros: conteúdos em texto completo e em formato digital prontos para serem disseminados;
- foco na comunidade: promovem a disseminação do conhecimento produzido pela instituição.

O RDI a ser analisado neste trabalho será o da UFPR. A UFPR é considerada a universidade mais antiga do Brasil e, de forma pioneira, criou seu RDI em 2004, que hoje conta com aproximadamente 50.000 objetos digitais em forma de teses, dissertações, vídeos, artigos de periódicos, e outros.

Vários estudos sobre criação e atualização de repositórios digitais institucionais orientando as boas práticas a serem seguidas são encontrados na literatura, conforme explorado nas pesquisas realizadas e que são apresentados no capítulo a seguir.

# 3 METADADOS, INTERPRETAÇÃO E INTEROPERABILIDADE

Metadados são informações vinculadas a um recurso armazenado, seja ele físico ou não. Metadados não só identificam e descrevem um objeto digital, como também documentam seu comportamento, função e uso, bem como sua relação com outros objetos digitais e como o mesmo deve ser gerenciado. Metadados são estruturados na forma de textos e palavras-chaves e suas informações, geralmente, são diretas, tais como nome do autor, data de criação, assunto, mas também podem ser complexas e mais difíceis de serem definidas, tal como o consenso da opinião de várias pessoas sobre um mesmo livro (LANGIANO, 2005).

Leite (2009), citando NISO[1], define metadados como "dados estruturados que descrevem, identificam, explicam, localizam e, portanto, facilitam a recuperação, uso e gestão de recursos de informação", de onde conclui-se sua importância para facilitar a descoberta de conteúdos relevantes em RDIs.

Objeto digital "é um objeto de informação, de qualquer tipo e formato expresso sob a forma digital" (YAMAOKA et al., 2013) tais como livros, capítulos de livros, periódicos, artigos e outros.

Um item ou objeto disponibilizado em meio digital deve sobreviver a gerações sucessivas de hardware e software. Considerando esta complexidade e a importância na concepção de metadados de objetos digitais, Baca (1998) propõe categorizá-los em 5 tipos, (Quadro 1), e posteriormente categorizá-los conforme suas várias características e funções, (Quadro 2).

Metadados contém informações passíveis de serem localizadas através de pesquisas o que auxilia na identificação e recuperação de determinados recursos de um objeto digital, por exemplo, um recurso digital visual que não possui nenhum texto que o descreva, é perfeitamente "pesquisável" pelas informações adicionadas a seus metadados, os chamados metadados descritivos, tais como direitos autorais, informações sobre o processo de criação de imagens, formato e etc, conforme exposto no Quadro 1 (LANGIANO, 2005).

[1] NATIONAL INFORMATION STANDARDS ORGANIZATION. **Understanding Metadata**. Bethesda, MD, EUA: NISO, 2004. ISBN: 1-880124-62-9.

QUADRO 1 – TIPOS DE METADADOS

| **Tipo** | **Definição** |
|---|---|
| Administrativo | utilizados na gestão e administração de recursos de informação; por exemplo, controle de versões e informações sobre direitos autorais. |
| Descritivo | usados para descrever e identificar informações sobre recursos; por exemplo, índices especializados e auxílios para busca. |
| Preservação | relacionam-se com a preservação de recursos de informação; por exemplo, políticas relacionadas a *backup* do objeto digital. |
| Técnico | relacionados com a operação ou o comportamento de metadados do sistema; por exemplo, processos de digitalização. |
| Uso | relacionados com o nível e tipo de utilização dos recursos de informação; por exemplo, estatísticas. |

FONTE: Adaptado de Baca (1998, p. 3).

A criação e gestão de metadados tornou-se um mistura muito complexa de processos manuais e automatizados criado por distintos indivíduos, com o intuito de promover a identificação correta e precisa de objetos digitais, assim sendo aos vários tipos de metadados e suas funções são adicionadas características e atributos que melhor identificam o objeto digital. O Quadro 2 demonstra alguns dos principais atributos por tipo de metadados com exemplos.

QUADRO 2 – ATRIBUTOS E CARACTERÍSTICAS DE METADADOS

| **Atributo** | **Características** |
|---|---|
| Fonte | – Metadados internos associados a um objeto digital no momento de sua criação ou digitalização.<br>– Metadados externos associados a um objeto digital posteriormente a sua criação, geralmente por alguém que não seja o agente inicial.<br>Exemplo: estruturas de diretório, formato de registro e esquema de compressão. |
| Método de Criação | – Metadados gerados automaticamente por computador.<br>– Metadados adicionados manualmente por usuários.<br>Exemplo: índices de palavras-chave. |
| Personagem | – Metadados adicionados, originalmente pelos autores do objeto digital, que muitas vezes não são especialistas em ciência da informação.<br>– Metadados adicionados por especialistas em ciência da informação.<br>Exemplo: sistemas pessoais de arquivamento, registros MARC. |
| Estado | – Metadados estáticos, que nunca mudam após a sua criação.<br>– Metadados dinâmicos, que podem mudar com o uso ou manipulação do objeto digital.<br>– Metadados que asseguram a longo prazo a disponibilidade do objeto digital.<br>– Metadados de curta duração, principalmente de tipo operacional.<br>Exemplo: informações sobre direitos autorais, conservação e administração da documentação. |
| Estrutura | – Metadados estruturados ou que seguem uma estrutura previamente padronizada.<br>– Metadados não estruturados ou que não seguem uma estrutura previamente padronizada.<br>Exemplo: formatos de bancos de dados locais, campos de notas não-estruturados. |
| Semântica | – Metadados controlados que seguem um vocabulário padrão ou uma forma de autoridade.<br>– Metadados não controlados ou que não seguem um vocabulário padrão ou uma forma de autoridade.<br>Exemplo: notas de texto livre, meta-*tags* HTML |
| Nível | – Metadados relacionados a coleções de objetos digitais.<br>– Metadados relacionados com objetos digitais individuais, muitas vezes incluídos dentro de uma coleção específica.<br>Exemplo: informações sobre formato, índices especializados. |

FONTE: Adaptado de Baca (1998, p. 4).

Este trabalho utilizará metadados descritivos, conforme Quadro 1, com qualquer atributo ou característica, de acordo com o Quadro 2, para identificação de conteúdos bibliográficos das produções científicas.

## 3.1 ESQUEMAS DE METADADOS

Esquemas de metadados são conjuntos de elementos projetados para um propósito especifico, ou seja, utilizados para descrever um recurso informacional. A definição ou significado dos elementos é conhecido como a semântica do esquema, e os valores de um dado elemento são os conteúdos. Os esquemas de metadados, geralmente, especificam os nomes dos elementos e as semânticas correspondentes (SAYÃO, 2007b).

Metadados devem ser cuidadosamente planejados e sempre obedecer um método que permita a interoperabilidade com outras instituições e, consequentemente, facilite a localização e uso do objeto digital. Esquemas e padrões de metadados existem para permitir o efetivo compartilhamento de recursos entre instituições e pessoas.

Abaixo, uma lista de alguns esquemas de metadados existentes e suas particularidades:

- *Machine-Readable Cataloguing* (MARC) – os formatos MARC são utilizados para a representação e comunicação de dados bibliográficos entre bibliotecas[2]. Os elementos de dados do MARC formam a base da maioria dos catálogos usados hoje em bibliotecas de todo o mundo.

- *Metadata Object Description Schema* (MODS) – é um subconjunto do esquema MARC 21 representado em formato XML, criado com a intenção de complementar outros esquemas de metadados.[3]

- *Encoded Archival Description* (EAD) – composto por um conjunto de regras para descrição biblioteconômica para permitir a catalogação de acervos em arquivos, museus, bibliotecas e similares.[4]

- *Learning Object Metadata* (LOM) – é um modelo de dados, codificado em XML, usado para descrever adequadamente objetos de aprendizagem.

- *Interoperability of Data in Commerce Systems* (Indecs) – é uma estrutura formal para identificação e descrição de propriedade intelectual e as entidades envolvidas.[5]

- *Dublin Core* (DC) – é um esquema de metadados, que surgiu em março de 1995 objetivando promover a interoperabilidade entre metadados (*ARXIV*, 2016). O *Dublin Core* utiliza um "conjunto de elementos simples, mas eficazes para descrever uma ampla variedade de recursos de rede" e "cuja semântica foi estabelecida por um

[2] LOC, 2015.
[3] 2014.
[4] 2012.
[5] INDECS, 1999.

consenso internacional de profissionais de diversas disciplinas, tais como biblioteconomia, computação, marcação de texto, a comunidade de museus e outras áreas afins" (LARA, 2015).

O Dublin Core foi o esquema de metadados adotado na presente dissertação pelo amplo uso na área de biblioteconomia e também por ser o esquema de metadados utilizado no RDI objeto de estudo deste trabalho.

Este esquema de metadados utiliza-se de quinze elementos descritivos conforme padronizado em vocabulários técnicos e especificações mantidas pela Iniciativa de Metadados Dublin Core (DCMI). A Iniciativa *Dublin Core* (*DC Metadata Initiative*) inclui também classes de recursos, esquemas de codificação e sintaxe de vocabulários, com o objetivo de promover o uso combinado dos termos DCMI com outros vocabulários compatíveis, considerando o contexto do perfil da aplicação e de acordo com o Modelo Abstrato DCMI (DCMI, 2012).

Os elementos deste esquema são identificados como "dc" e possuem valor único. Cada elemento tem ocorrência ilimitada, e para diferenciar o valor de cada ocorrência, são utilizados qualificadores, os quais podem ter um identificador, chamado esquema e/ou modificador (ALVES et al., 2007) conforme a sintaxe: **dc.*elemento.qualificador***. A Figura 1 traz um exemplo desta sintaxe.

FIGURA 1 – EXEMPLO DE ELEMENTOS *DUBLIN CORE*

| Elemento | Valor | Idioma |
|---|---|---|
| dc.contributor.author | Barbosa, Eduardo Mayer | |
| dc.contributor.author | Rodrigues, Tamires Maria | |
| dc.date.accessioned | 2015 06 13T00:31:47Z | |
| dc.date.available | 2015 06 13T00:31:47Z | |
| dc.date.issued | 2015 06 12 | |
| dc.identifier.uri | http://hdl.handle.net/1884/38213 | |
| dc.language.iso | pt_BR | pt_BR |
| dc.rights | Attribution 3.0 United States | * |
| dc.rights.uri | http://creativecommons.org/licenses/by/3.0/us/ | * |
| dc.subject | Esporte de Orientação, Leitura de Mapas, Geografia, Ensino. | pt_BR |
| dc.title | USO DO ESPORTE DE ORIENTAÇÃO EM AMBIENTE REDUZIDO PARA O ENSINO DE LEITURA DE MAPAS | pt_BR |
| dc.type | Working Paper | pt_BR |

FONTE: A autora (2015).

## 3.2 METADADOS E INTEROPERABILIDADE

As informações de objetos digitais armazenados em RDIs são chamadas de conteúdo e divididas em Dados e Metadados, onde "Dado" é o termo genérico para descrever as informações em formato digital e "Metadados" são dados sobre os dados (LANGIANO, 2005).

Metadados, se cuidadosamente construídos, trazem diversas vantagens para os usuários de Bibliotecas Digitais, pois por meio de uma representação padronizada dos recursos informacionais disponíveis em meio eletrônico, proporcionam o acesso amplo e preciso aos conteúdos nelas armazenados.

Como objetos digitais devem sobreviver a gerações sucessivas de hardware e software e também resistir às mudanças para novos sistemas, precisam possuir metadados que os permitam existir independentemente do sistema em uso para armazenamento e busca. Cabe salientar que para que os objetos digitais se mantenham acessíveis e inteligíveis ao longo do tempo é imprescindível que seja possível o transporte e preservação de seus metadados (BACA, 1998).

Para facilitar a busca e acesso aos conteúdos armazenados em RDIs é comum adotar um esquema de metadados para descrição de objetos digitais (vídeos, sons, imagens, textos e sites na web, etc) sempre de acordo com o objetivo do RDI.

Segundo (VALMORBIDA et al., 2011), em um estudo de dezembro de 2011, dentre os diversos protocolos para a interoperabilidade de metadados, é possível identificar duas vertentes genéricas e de cunho internacional relacionadas ao tema, uma delas voltada principalmente para a promoção de interoperabilidade de metadados de descrição de recursos digitais e outra voltada aos metadados de catalogação de documentos físicos. Neste mesmo estudo os autores apresentaram a implementação e análise de um sistema de banco integrado, considerando os principais protocolos para interoperabilidade de metadados existentes (VALMORBIDA et al., 2011) , destacando:

- O protocolo Z39.50, que especifica procedimentos e formatos para um cliente procurar um banco de dados fornecido por um servidor, recuperar os registros deste banco de dados, e executar funções de recuperação de informação relacionadas;

- Os protocolos SRU (Search/Retrieve URL) e SRW (Search/Retrieve Web Services), baseados no protocolo Z39.50,SRU, porém mais mais voltados às mudança advindas para a representação de registros catalográficos, até então binários para a linguagem XML (Extensible Markup Language) e;

- O protocolo OAI/PMH, que consiste na conversão dos metadados em um conjunto de elementos comuns que são disponibilizados por mecanismos de harvesting, sendo desta forma, colhidos e armazenados em bancos de dados, permitindo a busca dos metadados independentemente dos repositórios originais.

O protocolo OAI/PMH é o protocolo utilizado pela UFPR para promoção da interoperabilidade de seu RDI.

### 3.2.1 Protocolo OAI-PMH

Para promover a interoperabilidade entre RDIs é utilizado o protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

O OAI-PMH é um projeto da *Open Archives Initiative*, que surgiu para promover a interoperabilidade entre RDIs e tem suas raízes em um esforço para melhorar o acesso aos arquivos de *e-prints* como uma forma de aumentar a disponibilidade de comunicação acadêmica (OAI, 2016). Este protocolo define quais critérios devem ser observados para facilitar a disseminação eficiente de conteúdo nestes ambientes digitais. No conceito do OAI-PMH existem dois tipos de provedores:

- Os provedores de dados: que são repositórios que expõem seus metadados estruturados de acordo com o protocolo OAI-PMH e;
- Os provedores de serviços: que fazem solicitações de serviço via OAI-PMH para a colheita destes metadados expostos.

A adoção do OAI-PMH promove aumento da visibilidade e compartilhamento do conhecimento dos conteúdos de RDIs e de acordo com Oliveira (2010) "vem se consolidando como base para a interoperabilidade entre bibliotecas e repositórios digitais acadêmicos e científicos em todo o mundo." Um exemplo disso é a Biblioteca Digital Brasileira de Teses e Dissertações[6], que agrega teses e dissertações defendidas em todo o País e por brasileiros no exterior.

Embora algumas instituições adotem o OAI-PMH para colheita de artigos a serem inseridos em seus RDIs nem sempre é possível coletar a informação sob qual licença o artigo foi produzido, conforme exposto no capítulo 5. Para efeitos deste estudo foi criado um conjunto de componentes para extração e carga automática dos artigos científicos, produzidos pelos docentes da UFPR, já identificando-os quanto a forma de acesso utilizada na publicação, ou seja, se de acesso aberto ou não, a partir das bases de dados SHERPA/RoMEO, CrossRef e Plataforma Lattes.

Para a localização preciso de metadados em ambientes *web* é imprescindível a adoção dos chamados "identificadores persistentes", que são descritos no capitulo posterior.

## 3.3 OBJETOS DIGITAIS/IDENTIFICADORES PERSISTENTES

O arquivamento e preservação de materiais digitais a longo prazo é uma tarefa difícil e dispendiosa, que requer recursos substanciais e compromisso institucional (NISO, 2007). Em meados de 1990, com a popularização da *Web*, surgiram os identificadores persistentes que são elementos exclusivos de identificação adicionados a um objeto digital,

[6] Disponível em: http://bdtd.ibict.br

que independentemente de sua localização ou formato, garantem que o mesmo seja acessível a longo prazo, a despeito de mudanças físicas e tecnológicas (SAYÃO, 2007a).

### 3.3.1 Handle System

O identificador persistente *Handle System*® foi desenvolvido em 1994 pela Corporação para Iniciativas de Pesquisa Nacional (CNRI) nos Estados Unidos. É um componente para a arquitetura de objetos digitais, que fornece serviços de resolução seguros, eficientes e extensíveis para identificadores exclusivos e persistentes.

Resolução de serviços é o mecanismo que faz com que um determinado identificador persistente *link* para a *url* específica onde o objeto digital está armazenado.

Um *handle* é composto de um prefixo e um sufixo, separados por uma barra (`/`). O prefixo identifica as autoridades nomeadoras, que criam e mantém identificadores; já o sufixo identifica um objeto digital no domínio da autoridade nomeadora. Autoridades nomeadoras são definidas de forma hierárquica, sendo cada componente separado por um ponto (`.`).

### 3.3.2 DOI

O Digital Object Identifier (DOI) foi apresentado pela primeira vez no *Frankfurt Book Fair* em 1997 e neste mesmo ano foi criada a *International DOI Foundation* (IDF) para gerenciar este sistema.

O DOI que é uma implementação proprietária do *Handle System*, originou-se de uma iniciativa conjunta de três associações comerciais na indústria editorial (*International Publishers Association*; *International Association of Scientific, Technical and Medical Publishers*; e *Association of American Publishers*) como uma estrutura genérica para a gestão de identificação de conteúdo através de redes digitais (DOI, 2015). Desde então, DOIs são utilizados para atribuir e disseminar informações sobre direitos de propriedade intelectual a objetos digitais (SAYÃO, 2007a).

A sintaxe de um identificador DOI é similar à de um *handle* e é formado por dois elementos: um prefixo fornecido por uma autoridade nomeadora, atualmente sempre 10, seguido de um código atribuído pela IDF a grupos como editores, publicadores e detentores de direitos autorais, os chamados gestores de conteúdo; e um sufixo atribuído pelo registrante a um objeto específico. (SAYÃO, 2007a)

Para a correta localização de um objeto digital através de um DOI, é necessário que o mesmo possua um mínimo de metadados estruturados, tais como informações bibliográficas e comerciais. Metadados atribuídos a um objeto digital dão ao usuário a garantia de que o recurso encontrado é efetivamente o que estava sendo procurado.

O modelo de dados utilizado por um identificador DOI provê um sistema contextual de metadados que suportam a interoperabilidade entre os diversos esquemas de metadados existentes em um ambiente digital. Este modelo consiste de um dicionário de dados interoperável acrescido de uma estrutura subjacente para aplicação.

Para atribuição de um DOI a um objeto digital com o intuito de permitir sua localização de forma única e persistente em um ambiente *WEB* é necessário que seus metadados sejam depositados previamente pelos seus publicadores científicos no sistema CrossRef (FERREIRA et al., 2015).

Este trabalho utiliza o DOI, para busca na Plataforma Lattes e no *site* https://doi.org/, por ser este o identificador único permanente mais comumente utilizado para a busca e recuperação de artigos científicos no ambiente online, como RDIs, que são explanados no capítulo 2.

# 4 IDENTIFICADORES PERSISTENTES

O arquivamento e preservação de materiais digitais a longo prazo é uma tarefa difícil e dispendiosa, que requer recursos substanciais e compromisso institucional (NISO, 2007). Em meados de 1990, com a popularização da *Web*, surgiram os identificadores persistentes que são elementos exclusivos de identificação adicionados a um objeto digital, que independentemente de sua localização ou formato, garantem que o mesmo seja acessível a longo prazo, a despeito de mudanças físicas e tecnológicas (SAYÃO, 2007a).

## 4.1 HANDLE SYSTEM

O identificador persistente *Handle System*® foi desenvolvido em 1994 pela CNRI nos Estados Unidos. É um componente para a arquitetura de objetos digitais, que fornece serviços de resolução seguros, eficientes e extensíveis para identificadores exclusivos e persistentes (CNRI, 2015).

Resolução de serviços é o mecanismo que faz com que um determinado identificador persistente *link* para a *url* específica onde o objeto digital está armazenado.

Um *handle* é composto de um prefixo e um sufixo, separados por uma barra (/). O prefixo identifica as autoridades nomeadoras, que criam e mantém identificadores; já o sufixo identifica um objeto digital no domínio da autoridade nomeadora. Autoridades nomeadoras são definidas de forma hierárquica, sendo cada componente separado por um ponto (.).

## 4.2 DOI

O DOI foi apresentado pela primeira vez no *Frankfurt Book Fair* em 1997 e neste mesmo ano foi criada a IDF para gerenciar este sistema.

O DOI que é uma implementação proprietária do *Handle System*, originou-se de uma iniciativa conjunta de três associações comerciais na indústria editorial (*International Publishers Association*; *International Association of Scientific, Technical and Medical Publishers*; e *Association of American Publishers*) como uma estrutura genérica para a gestão de identificação de conteúdo através de redes digitais (DOI, 2015). Desde então, DOIs são utilizados para atribuir e disseminar informações sobre direitos de propriedade intelectual a objetos digitais (SAYÃO, 2007a).

A sintaxe de um identificador DOI é similar à de um *handle* e é formado por dois elementos: um prefixo fornecido por uma autoridade nomeadora, atualmente sempre 10, seguido de um código atribuído pela IDF a grupos como editores, publicadores e detentores de direitos autorais, os chamados gestores de conteúdo; e um sufixo atribuído pelo registrante a um objeto específico. (SAYÃO, 2007a)

Para a correta localização de um objeto digital através de um DOI, é necessário que o mesmo possua um mínimo de metadados estruturados, tais como informações bibliográficas e comerciais. Metadados atribuídos a um objeto digital dão ao usuário a garantia de que o recurso encontrado é efetivamente o que estava sendo procurado.

O modelo de dados utilizado por um identificador DOI provê um sistema contextual de metadados que suportam a interoperabilidade entre os diversos esquemas de metadados existentes em um ambiente digital. Este modelo consiste de um dicionário de dados interoperável acrescido de uma estrutura subjacente para aplicação.

Para atribuição de um DOI a um objeto digital com o intuito de permitir sua localização de forma única e persistente em um ambiente *WEB* é necessário que seus metadados sejam depositados previamente pelos seus publicadores científicos no sistema CrossRef (FERREIRA et al., 2015).

Este trabalho utiliza o DOI, para busca na Plataforma Lattes e no *site* https://doi.org/, por ser este o identificador único permanente mais comumente utilizado para a busca e recuperação de artigos científicos no ambiente online, como RDIs, que são explanados no próximo capítulo.

# 5 PANORAMA GERAL DA COLHEITA DE METADADOS NAS INSTITUIÇÕES

Por promover e contribuir para disseminação da produção científica os repositórios digitais institucionais devem estar sempre disponíveis e serem constantemente atualizados, pois também fazem parte das ferramentas que garantem a visibilidade da instituição e de seus pesquisadores.

É mister para o sucesso de um RDI a interação efetiva entre as equipes de desenvolvimento e manutenção das ferramentas tecnológicas e a equipe responsável pelo acervo.

Analisando-se rankings universitários, que são "listas de instituições de ensino superior, ordenadas usando uma combinação de indicadores" (MAHASSEN, 2014, p. 1), identificou-se que um dos indicadores utilizados para tal classificação é o volume de produção científica constante em RDIs e, dos 1000 RDIs classificados por um destes rankings, no caso o Webometrics (2014), setecentas e setenta e uma instituições internacionais e dezesseis instituições brasileiras apresentam seus artigos científicos agregados dentro de seu RDI ou em um repositório específico como por exemplo, a Universidade de São Paulo.

Além dos rankings universitários, os RDIs, também podem aumentar o impacto das publicações científicas de uma instituição.

RDIs podem ser atualizados pela colheita de metadados obtidos através do uso do OAI-PMH, lembrando que para isso é necessário a configuração prévia dos recursos de interoperabilidade do referido protocolo (WADHAM, 2002), ou ainda pelo processo de auto-submissão, onde os próprios autores depositam sua produção no RDI desejado.

## 5.1 COLHEITA DE METADADOS POR OAI-PMH

O uso do protocolo OAI-PMH para colheita automática de metadados em bibliotecas digitais baseia-se na extração de metadados de bases de dados bibliográficas, tais como *Scopus*[1], *Web of Science*[2], *SciELO*[3] e outras, e enfrenta, além da falta de padronização adotada para construção de metadados de publicações científicas disponibilizadas sob uma licença de acesso aberto, também a dificuldade de que artigos científicos, em acesso aberto ou não, só são encontrados nestas bases de dados se atenderem às diretrizes específicas das mesmas, o que acaba por torná-lo não efetivo para localização de todos os artigos produzidos por uma instituição.

[1] Disponível em: http://www.scopus.com/
[2] Disponível em: http://webofscience.com/
[3] Disponível em: http://www.scielo.org/php/index.php

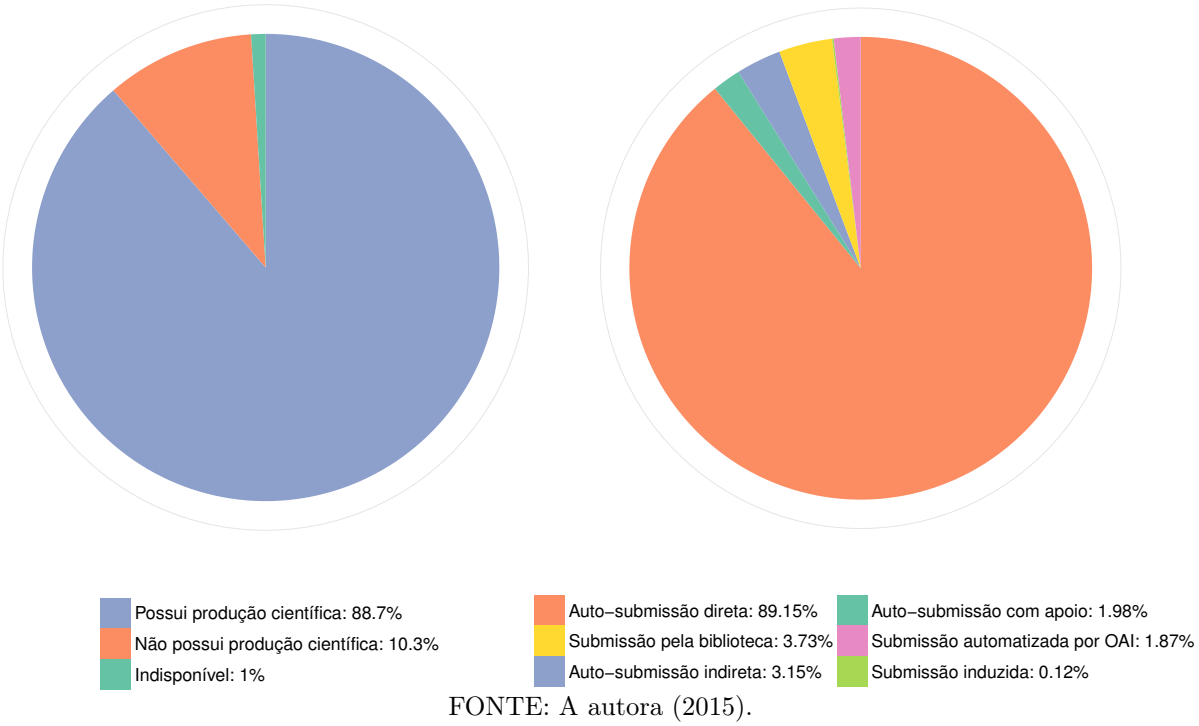
## 5.2 INCLUSÃO DE ARTIGOS CIENTÍFICOS POR AUTO-SUBMISSÃO

Auto-arquivamento ou auto-submissão é o processo de publicação da produção científica em RDIs pelo próprio autor sem intermédio de terceiros.

O armazenamento de artigos científicos em RDIs, identificados neste estudo, utiliza-se do processo de auto-submissão para colheita da produção científica. A auto-submissão recebe, em alguns casos específicos, apoio de equipe específica da área de biblioteconomia ou de ferramentas customizadas para recuperação de metadados, além da participação indispensável dos autores para informações quanto ao licenciamento desta produção científica.

Dos 1000 primeiros repositórios institucionais classificados pelo ranking *Webometrics* no segundo semestre do ano de 2014, conforme mostrado na Figura 2, 887 possuem artigos científicos, 103 possuem apenas monografias, teses e dissertações e 10 estavam indisponíveis. Dos 887 que possuem artigos científicos, 764 utilizam o processo de auto-submissão direta, 17 auto-submissão com apoio de ferramentas para recuperação de metadados, 32 a submissão é realizada pela biblioteca após recebimento dos dados e metadados enviados pelos autores, 27 a auto-submissão é feita indiretamente pelos autores com apoio de uma equipe da área de biblioteconomia, 16 possuem colheita automatizada por OAI-PMH e em 1 deles uma equipe da biblioteca identifica a produção científica a partir do endereço de correio eletrônico de um dos autores e, posteriormente solicita ao mesmo que se cadastre e submeta a publicação na biblioteca digital da instituição (submissão induzida).

FIGURA 2 – REPOSITÓRIOS X PRODUÇÃO DE ARTIGOS CIENTÍFICOS




FONTE: A autora (2015).

Alguns dos repositórios analisados são povoados pela obtenção conjunta do processo de auto-sub-missão mais a colheita por OAI/PMH em bases institucionais ou internacionais, como por exemplificado nos parágrafos abaixo.

Os repositórios das instituições suecas, *Uppsala University*[4], *Linköping University*[5], *Umea University*[6], *Stockhom University*[7], *Royal Institute of Technology*[8] e o da norueguesa *Norwegian University of Science and Technology*[9] utilizam a opção de auto-submissão mais colheita incremental à base Digitala Vetenskapliga Arkivet (DiVA). DiVA é uma ferramenta de descoberta para artigos científicos e teses de estudantes com 37 universidades participantes (DIVA, 2015). Os artigos científicos das universidades participantes são obtidos por auto-arquivo e também por colheita automatizada das bases de dados *Web of Science*[10], *Scopus*[11] e *PubMed*[12].

Nos repositórios das universidades irlandesas, *National University of Ireland Galway* (ARAN, 2015) e *University College Cork* (CORA, 2015) os artigos científicos são obtidos por colheita automatizada em seus "Sistema de informação de pesquisa institucional" (IRIS). Os sistemas IRIS das duas instituições são alimentados pelo processo de auto-arquivo.

Os repositórios holandeses da *Universiteit Utrecht*[13] e da *Vrije Universiteit Amsterdam*[14] obtém os artigos científicos por colheita automatizada na base de dados holandesa NARCIS (2015), que é uma iniciativa das universidades holandesas para publicação de seus resultados de pesquisas em acesso aberto.

A *University of Portsmouth*[15], no Reino Unido, e a *University of St Andrews*[16] na Escócia fazem colheita automatizada de artigos científicos em seus sistemas de gerenciamento de informação denominados *Pure* [17], que são alimentados por auto-arquivo pelos autores/pesquisadores da instituição.

A *Vietnam National University*[18] atualiza seus artigos científicos por colheita incremental nas bases de dados *Scopus*[19] e *Journal of Science*[20].

4 Disponível em: http://uu.diva-portal.org/
5 Disponível em: http://liu.diva-portal.org/
6 Disponível em: http://umu.diva-portal.org/
7 Disponível em: http://su.diva-portal.org/
8 Disponível em: http://kth.diva-portal.org/
9 Disponível em: http://ntnu.diva-portal.org/
10 Disponível em: http://webofscience.com/
11 Disponível em: http://www.scopus.com/
12 Disponível em: http://www.ncbi.nlm.nih.gov/pubmed
13 Disponível em: http://igitur-archive.library.uu.nl/
14 Disponível em: http://dare.ubvu.vu.nl/
15 Disponível em: http://eprints.port.ac.uk/
16 Disponível em: https://research-repository.st-andrews.ac.uk/
17 *Pure* é uma solução para gerenciamento de produção científica fornecida pela editora Elsevier.
18 Disponível em: http://dl.vnu.edu.vn/
19 Disponível em: http://www.scopus.com/
20 Disponível em: http://www.sciencemag.org/journals

O "Repositório Aberto da Universidade do Porto" em Portugal[21] possui integração para colheita automatizada de seus artigos científicos que são armazenados por auto-arquivo em seu "Sistema de Informação para Gestão Agregada dos Recursos e dos Registros Acadêmicos" (Sigarra).

A *City University London* possui uma ferramenta, denominada CRIS (*Current Research Information System*), que varre ambientes *web* utilizando como filtro de pesquisa os endereços de e-mail de seus pesquisadores/autores no intuito de localizar metadados de publicações científicas a estes relacionados e, logo em seguida encaminha aos selecionados solicitação de inclusão destas publicações em seu repositório por auto-arquivo (CITY RESEARCH ONLINE, 2015).

Existem 23 instituições brasileiras que possuem repositórios digitais institucionais classificados entre os top 1000 repositórios do ranking *Webometrics*, conforme demonstrado na Figura 3, sendo que destas 16 possuem artigos científicos em seu repositório e apenas duas: a universidade São Paulo e a Universidade do Rio Grande do sul possuem ferramentas específicas para gerenciamento da produção científica institucional, respectivamente os sistemas SABI e Dédalus.

FIGURA 3 – POSIÇÃO DOS REPOSITÓRIOS BRASILEIROS NO RANKING WEBOMETRICS

| BRASIL | WORLD | REPOSITÓRIO | ART? |
|---|---|---|---|
| 1 | 11 | Repositório Digital Universidade Federal do Rio Grande do Sul LUME | sim |
| 2 | 13 | Universidade de São Paulo Biblioteca Digital de Teses e Dissertações | não |
| 3 | 69 | Universidade Federal do Parana Biblioteca Digital de Teses e Dissertações | não |
| 4 | 89 | Repositório Institucional Universidade Federal de Santa Catarina | não |
| 5 | 162 | Universidade Federal da Bahia Repositorio Institucional | sim |
| 6 | 193 | Universidade de Brasília Repository | sim |
| 7 | 207 | Alice Repository Open Access to Scientific Information Embrapa | sim |
| 8 | 432 | Repositório Institucional UNESP Universidade Estadual Paulista Júlio de Mesquita Filho | sim |
| 9 | 515 | Repositorio Institucional Universidade Federal do Ceará | sim |
| 10 | 571 | Acervo Digital da Universidade Estadual Paulista Júlio de Mesquita Filho | sim |
| 11 | 591 | Repositorio Institucional Fundação Oswaldo Cruz | sim |
| 12 | 618 | Repositório Institucional Universidade Federal do Rio Grande | sim |
| 13 | 700 | Biblioteca Digital da Produção Intelectual da Universidade de São Paulo | sim |
| 14 | 710 | Repositório Institucional Universidade Federal do Rio Grande do Norte | sim |
| 15 | 719 | Repositório Institucional Universidade Federal do Pará | sim |
| 16 | 771 | Repositório Institucional Centro Universitário de Brasília | não |
| 17 | 797 | Maxwell Pontifícia Universidade Católica do Rio de Janeiro | não |
| 18 | 798 | Repositório Institucional Universidade Federal de Lavras | sim |
| 19 | 832 | Repositorio Institucional Pontificia Universidad Católica de Río Grande do Sul | sim |
| 20 | 936 | Repositório Institucional Universidade Federal de Goiás | sim |
| 21 | 948 | Repositório de Outras Coleções Abertas Universidade Tecnológica Federal do Paraná ROCA | não |
| 22 | 965 | Biblioteca Digital de Monografias de Graduação e Especialização Universidade de Brasília | não |
| 23 | 990 | CBPF Index Centro Brasileiro de Pesquisas Físicas | sim |

FONTE: Adaptado de Webometrics (2015).

Na Universidade Federal do Rio Grande do Sul as produções científicas são obtidas por uma força-tarefa que envolve docentes, alunos, equipe de desenvolvimento e responsáveis pelo acervo na correta alimentação do seu Sistema de Automação de Bibliotecas (SABi) e posterior carga automatizada no repositório da instituição por colheita incremental (PAVÃO; COSTA et al., 2013).

[21] Disponível em: http://repositorio-aberto.up.pt/

A UFRGS tem uma política de acompanhamento e registro da produção intelectual de seus membros que é anterior à existência dos RDIs. Daí decorre seu sucesso.

A carga inicial dos artigos científicos na Universidade de São Paulo no ano de 2012 foi realizada de forma manual e após 2013 incluiu a importação de dados coletados da base *SciELO*[22], além da opção de auto-depósito por docentes e pós-graduandos. (MURAKAMI, 2015). Atualmente, a equipe de informática da Universidade está trabalhando no desenvolvimento de uma ferramenta para coleta e importação de dados do Banco Bibliográfico de Dados (Dedalus) através do protocolo OAI-PMH, solução esta similar à adotada pela Universidade do Rio Grande do Sul.

Ramos et al. (2012) apresentaram um estudo, utilizando o OAI-PMH, e as devidas limitações existentes para carga automática em um repositório digital de registros contidos em bases de dados internacionais. Os autores acreditam, como de fato hoje se comprova, que

> [...] "a carga automática de registros referenciais tem muito a colaborar com o início do desenvolvimento dos repositórios, à medida que oferece visibilidade à produção cientifica, favorecendo as instituições na elaboração dos índices de produção científica extraídos a partir dos emergentes estudos webmétricos, sem desrespeitar os princípios de direitos do autor ou das editoras." (RAMOS et al., 2012, p. 91)

O acesso à informação científica produzido pelas instituições de ensino, com o uso adequado das tecnologias de informação e comunicação, em repositórios institucionais de acesso aberto, é apontado por Pavão, Sousa et al. (2009) como um dos elementos mais enfatizados para possibilitar a alteração do *status quo* e consequente inclusão social de indivíduos, pertencentes ou não a círculos acadêmicos, levando-se em conta a atual configuração do capitalismo global nas sociedades periféricas, o que é reforçado por Meadows (1999), que afirma ser um dos fatores importantes na divulgação de pesquisas a publicação efetiva de seus resultados.

Os estudos focam-se em carga ou atualização automática de repositórios baseados na extração de metadados de bases de dados nacionais e internacionais ou periódicos específicos.

Além da falta de padronização adotada para construção de metadados, já sinalizada, existe o desafio de como identificar publicações científicas produzidas e disponibilizadas em acesso aberto apenas a partir do periódico na qual a mesma está armazenada, porque conforme já mencionado, os artigos científicos, em acesso aberto ou não, só são encontrados em bases de dados bibliográficas se atenderem à diretrizes específicas. Da mesma forma, que essas bases não indexam toda a produção científica de uma instituição,

[22] Disponível em: http://www.scielo.org/php/index.php

nem toda a produção dela é disponibilizada em acesso aberto, essas bases não são efetivas para localização de todos os artigos produzidos por uma instituição, mesmo com a utilização do OAI-PMH para colheita de metadados e artigos científicos.

# 6 BASES DE METADADOS

Na literatura revisada encontram-se estudos para carga automatizada de repositórios digitais, porém partem sempre da importação de metadados constantes em bases de metadados normalizadas onde também são encontradas informações acerca do licenciamento da produção científica. Essas bases de metadados são atualizadas através do envolvimento dos docentes e pesquisadores e pelos responsáveis pelo acervo.

Em todos os repositórios digitais institucionais analisados que utilizam a opção de auto-arquivo direto, fica a cargo do pesquisador/autor a responsabilidade quanto à verificação do licenciamento de sua produção científica.

Apesar de alguns repositórios oferecerem ferramentas para facilitar a execução do processo de auto-submissão pelos pesquisadores/autores, é inexistente em qualquer um dos repositórios analisados e em qualquer trabalho científico a possibilidade da importação automática dos metadados, resumos e afins, sem que seja sempre necessário o envolvimento do pesquisador para completar as informações necessárias para importação nos repositórios. Oferece-se nestes repositórios, para facilitar esta atividade pelo pesquisador, *links* para ferramentas como o Directory of Open Access Journals (DOAJ), SHERPA/RoMEO, Diadorim, Dulcinea e outras, que indicam se um periódico foi disponibilizado em acesso aberto ou não. Ocorre porém que estas iniciativas não contemplam todos os periódicos existentes, sendo necessário, em muitos casos, que o autor/pesquisador contate o editor do periódico para saber acerca deste processo.

Como a partir de um identificador DOI obtém-se os metadados relacionados ao objeto digital, as bases de dados bibliográficas, tais como *Scopus*[1], *Web of Science*[2] e outras, o utilizam para recuperação no ambiente *online*.

## 6.1 CROSSREF

A CrossRef é uma associação de editores científicos, fundada no ano 2000, com o objetivo de promover referência cruzada entre publicações científicas. Para tanto é mantido um banco de metadados que referencia o conteúdo propriamente dito que permanece armazenado no *website* dos publicadores científicos. É também uma Agência Registradora para o sistema DOI orientada para cobertura de produções científicas e conta, atualmente, com mais de 75 milhões de DOIs registrados.

Para captura de metadados nela armazenados, a CrossRef fornece diversos Application Programming Interfaces (APIs) para atendimento à finalidades diversas. Para utilização de certos APIs é necessário o registro junto à mesma.

[1] Disponível em: http://www.scopus.com/
[2] Disponível em: http://webofscience.com/

## 6.2 A PLATAFORMA LATTES

A Plataforma Lattes (2012), "representa a experiência do Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações".

Atualmente, os docentes e pesquisadores de instituições brasileiras que produzem conhecimento científico e participam de programas como CAPES e CNPq são orientados a informarem suas produções científicas na Plataforma Lattes, logo a partir da mesma é possível a uma instituição de ensino acessar a produção científica de seu corpo docente através do "Sistema Lattes Extrator".

A extração de dados é fornecida por meio de arquivo XML que contém toda a produção científica da instituição por grupos de pesquisa, professores, pesquisadores e alunos registrados na plataforma.

As Figuras 4 e 5 abaixo demonstram um arquivo XML gerado e sua estrutura:

FIGURA 4 – EXEMPLO DE EXTRAÇÃO DE DADOS A PARTIR DA PLATAFORMA LATTES


FONTE: Plataforma Lattes (2012)

Uma informação imprescindível para o desenvolvimento do presente estudo é o DOI dos artigos publicados, presente no atributo de mesmo nome constante do arquivo XML disponibilizado pela Plataforma Lattes, conforme Figura 5.

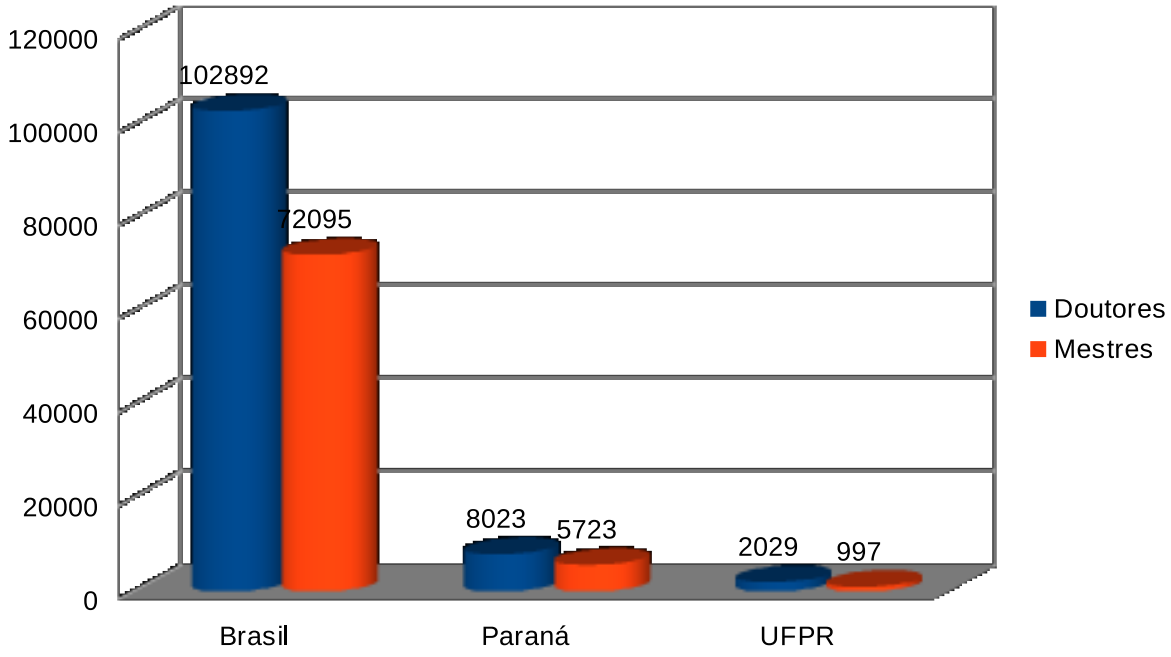
FIGURA 5 – IDENTIFICAÇÃO DO ARTIGO POR DOI

```
<ARTIGOS-PUBLICADOS>
        <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="146" ORDEM-IMPORTANCIA="">
                <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Extending OAI-PMH over
structured P2P networks for digital preservation" ANO-DO-ARTIGO="2012" PAIS-DE-PUBLICACAO=""
IDIOMA="Inglês" MEIO-DE-DIVULGACAO="MEIO_DIGITAL" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="SIM"
DOI="10.1007/s00799-012-0080-5" TITULO-DO-ARTIGO-INGLES="Extending OAI-PMH over structured P2P
networks for digital preservation" FLAG-DIVULGACAO-CIENTIFICA="NAO"/>
                <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="International Journal on
Digital Libraries (Print)" ISSN="14325012" VOLUME="12" FASCICULO="" SERIE="1" PAGINA-INICIAL="1"
PAGINA-FINAL="14" LOCAL-DE-PUBLICACAO=""/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Seára, Everton F. R." NOME-PARA-CITACAO="Seára,
Everton F. R." ORDEM-DE-AUTORIA="1" NRO-ID-CNPQ=""/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Bona, Luis C. E." NOME-PARA-CITACAO="Bona, Luis C.
E." ORDEM-DE-AUTORIA="3" NRO-ID-CNPQ="9945289519054859"/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Vignatti, Tiago" NOME-PARA-CITACAO="Vignatti, Tiago"
ORDEM-DE-AUTORIA="4" NRO-ID-CNPQ=""/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Vignatti, Andre L." NOME-PARA-CITACAO="Vignatti,
Andre L." ORDEM-DE-AUTORIA="5" NRO-ID-CNPQ=""/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Doucet, Anne" NOME-PARA-CITACAO="Doucet, Anne" ORDEM-
DE-AUTORIA="6" NRO-ID-CNPQ=""/>
                <AUTORES NOME-COMPLETO-DO-AUTOR="Marcos Sfair Sunye" NOME-PARA-CITACAO="SUNYE, Marcos
Sfair;Sunye, Marcos S." ORDEM-DE-AUTORIA="2"/>
        </ARTIGO-PUBLICADO>
</ARTIGOS-PUBLICADOS>
```

FONTE: A autora (2015).

A Figura 6 demonstra o número atual de doutores e mestres cadastrados na Plataforma Lattes pertencentes à UFPR, os dados demonstrados foram extraídos da Plataforma pelo serviço *Lattestats*, que fornece informações classificadas por instituição, região sexo e idade constantes dos currículos dos docentes na referida plataforma (PLATAFORMA LATTES, 2016).

FIGURA 6 – NÚMERO DE DOCENTES NO BRASIL, NO PARANÁ E NA UFPR POR TITULAÇÃO




FONTE: Adaptado de Plataforma Lattes (2015).
NOTA: Dados relativos à extração de dados da Plataforma Lattes em 31/01/2015.

# 7 DIREITOS AUTORAIS E ACESSO ABERTO

*Copyright* ou direito do autor é um termo jurídico usado para descrever os direitos que autores têm sobre suas obras artísticas e literárias.

## 7.1 DIREITOS AUTORAIS

Os direitos autorais são agrupados em direitos patrimoniais, relacionados à exploração financeira da obra, e direitos morais, que protegem os interesses não patrimoniais do autor (WIPO, 2015).

As leis de direitos autorais diferem de país para país e a maioria afirma que o proprietário dos direitos tem o direito de autorizar ou impedir certos usos em relação à uma obra ou, em alguns casos, receber remuneração pelo uso de seu trabalho (como através de uma gestão coletiva). O titular dos direitos patrimoniais de um trabalho pode proibir ou autorizar sua reprodução sob várias formas seja a publicação impressa ou gravação de som; a apresentação pública, como por exemplo, um jogo de futebol ou show musical; a cópia, por exemplo, na forma de discos ou DVDs; a veiculação, por rádio, cabo e satélite; a tradução para outros idiomas e a adaptação, como por exemplo um romance em um roteiro de cinema conforme os artigos oitavo, nono, décimos primeiro, segundo e quarto da Convenção de Berna para a Proteção das Obras Literárias e Artísticas (WIPO, 1971). Exemplos de direitos morais reconhecidos incluem o direito de reivindicar a autoria de uma obra e o direito de opor-se a alterações em um trabalho que podem vir a prejudicar a reputação do criador (WIPO, 2015).

Jungmann et al. (2010) define que, no âmbito do direito autoral, o titular do bem poderá permitir aos interessados o uso e distribuição de suas obras de acordo com condições específicas. Por exemplo, enquanto obras disponibilizadas sob *copyright* possuem uma única condição "não dão liberdade alguma" quanto à sua utilização, outras formas de licenciamento mantém protegidos os direitos do autor, mas permitem remixagem, distribuição e etc desde que observadas os termos de licenciamento.

Proposto por Richard Stallman, então pesquisador do MIT, o termo *copyleft* é um trocadilho com *copyright* e designa licenças que permitem que as obras licenciadas sejam executadas, modificadas e distribuídas livremente desde que as cópias e derivações mantenham a mesma forma de licenciamento. Embora originalmente aplicado a programas de computador, "este conceito se expandiu ao longo do tempo e atualmente é usado também para publicações" conforme Setenareski (2013, p. 48).

## 7.2 ACESSO ABERTO

Conforme a *Budapest Open Access Initiative* (2002), acesso aberto é "o acesso totalmente irrestrito e gratuito por parte de qualquer cientista, acadêmico, professor, estudante ou outro interessado" a publicações acadêmicas ou científicas. Para melhor ilustrar esta definição, Harnad (2015) estabeleceu os termos 'via dourada' relacionado à produção de periódicos científicos em acesso aberto e 'via verde' relacionada à produção científica disponibilizada em repositórios digitais institucionais de acesso aberto.

O acesso aberto amplia a difusão e aproveitamento de conhecimento e confere às obras e seus autores maior visibilidade e impacto por meio de citações que, de acordo com a ABNT (2002), são "a transcrição textual literal de parte da obra do autor consultado". Citações são a forma mais usual de medir o impacto da literatura científica que é a base do mundo acadêmico, tais como livros, capítulos de livros, periódicos, artigos e outros.

A *Berlin Declaration on Open Access* (2003) define que para o estabelecimento do acesso livre vantajoso, é necessário "o empenho ativo de todo e qualquer indivíduo que produza conhecimento científico ou seja detentor de patrimônio cultural". Ainda conforme a mesma Declaração (*BERLIN DECLARATION ON OPEN ACCESS*, 2003) é assegurado à comunidade a livre utilização e distribuição respeitando-se a correta atribuição da autoria.

É importante salientar que "acesso aberto" não significa que a publicação está disponível em domínio público, mas sim que está sujeita a uma licença que, embora mais permissiva que o *copyleft* e o *copyright*, reserva ao autor o direito de decidir como a obra pode ser utilizada. Por exemplo, no Brasil, quando uma obra entra em domínio público, mesmo que tenha sido atrelada a uma licença copyright, torna-se livre para qualquer forma de utilização respeitados os direitos morais do autor que são inalienáveis e irrenunciáveis (BRASIL, 1998, art. 27).

## 7.3 INICIATIVAS PARA IDENTIFICAR PRODUÇÃO EM ACESSO ABERTO

Para facilitar a identificação de periódicos científicos de acesso aberto e suas políticas, existem algumas iniciativas que se propõem a catalogar tais informações, desde que em conformidade com suas diretrizes, e mediante cadastro prévio por parte dos próprios periódicos, tais como:

- DOAJ[1] — indexa periódicos revisados por pares e de acesso aberto;
- Diadorim[2] — lista políticas de acesso de periódicos brasileiros de acesso aberto.
- Dulcinea[3] — lista políticas de acesso de periódicos espanhóis de acesso aberto;

[1] DOAJ, 2014.
[2] DIADORIM, 2014.
[3] DULCINEA, 2014.

- SHERPA/RoMEO[4] — lista políticas de acesso de periódicos

O RoMEO é parte dos Serviços SHERPA da University of Nottingham e funciona a partir da colaboração com outros parceiros por exemplo, o DOAJ e o National Center for Biotechnology Information (NCBI).

O SHERPA/RoMEO foi escolhido pelo presente estudo por ser um banco de dados que permite localizar políticas de editores em relação à auto-arquivamento de artigos de revistas na *web* e em repositórios de acesso aberto. O SHERPA/RoMEO fornece listagens, seguindo as definições aceitas para efeitos de auto-arquivo, principalmente, para uso pela comunidade acadêmica, conforme exemplo mostrado na Figura 7.

FIGURA 7 – FRAGMENTO DE RESULTADO DE PESQUISA NO SHERPA/ROMEO

```
<!-- [...] -->
<romeoapi version="2.9.9">
  <!-- [...] -->
  <journals>
    <journal>
      <jtitle>Physical Review A</jtitle>
      <issn>1050-2947</issn>
      <zetocpub>American Physical Society</zetocpub>
      <romeopub>American Physical Society</romeopub>
    </journal>
  </journals>
  <publishers>
    <publisher id="10">
      <name>American Physical Society</name>
      <!-- [...] -->
      <preprints>
        <prearchiving>can</prearchiving>
        <prerestrictions />
      </preprints>
      <postprints>
        <postarchiving>can</postarchiving>
        <postrestrictions />
      </postprints>
      <pdfversion>
        <pdfarchiving>can</pdfarchiving>
        <pdfrestrictions />
      </pdfversion>
      <!-- [...] -->
      <romeocolour>green</romeocolour>
    </publisher>
  </publishers>
</romeoapi>
```

FONTE: Adaptado de SHERPA/RoMEO (2015a).

### 7.3.1 Licenças de acesso aberto

Licenciar é ceder a outrem por tempo indeterminado e de forma não exclusiva o direito de uso de uma propriedade intelectual (ROCHA, 2007).

[4] SHERPA/ROMEO, 2014.

Atualmente existem duas formas de licenciamento em acesso aberto, as licenças STM (Science, Technology and Medical Publishers) e as licenças Creative Commons.

#### 7.3.1.1 Licenças-modelo STM

Com o advento do acesso aberto, alguns autores, editores e sociedades preocupados com o uso comercial de artigos de periódicos de segmentos específicos, como medicina por exemplo, sugeriram, em 2014, as chamadas licenças de acesso aberto STM (Science, Technology and Medical Publishers). Essas licenças clarificam as condições quanto ao uso comercial e concessões quanto à liberdade de tradução e mineração de textos e dados, ou seja, é possível traduzir a obra desde que observados determinados aspectos previamente contidos na licença (ISTM, 2015).

Licenças STM foram criticadas por limitar o uso, reuso e exploração de pesquisas de um segmento específico em contraponto ao objetivo do movimento Acesso Aberto que é fomentar a distribuição de conhecimento em benefício da ciência e sociedade (PLOS ADVOCACY, 2014).

#### 7.3.1.2 Creative Commons

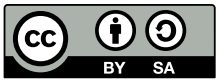
As licenças *Creative Commons* são instrumentos que fornecem a criadores individuais e até a grandes empresas, uma forma padronizada de atribuir autorizações de direito de autor e de direitos conexos aos seus trabalhos criativos ao mesmo tempo que permitem que outros copiem, distribuam e façam uso de seu trabalho de acordo com as opções permitidas pelo criador.

As licenças *Creative Commons* foram desenvolvidas pela organização sem fins lucrativos de mesmo nome, fundada em 2001 para possibilitar o livre compartilhamento e uso da criatividade e do conhecimento através de ferramentas legais (CREATIVE COMMONS, 2015). Em 2009, estimava-se que mais de 350 milhões de obras já haviam sido licenciadas sob uma licença *Creative Commons*.

As opções previstas pelas licenças *Creative Commons* são expressas através de símbolos que, quando combinados, dão forma às várias licenças, sendo que a única condição comum a todas elas é o reconhecimento do autor original conforme demonstrado pela Figura 8:

FIGURA 8 – SÍMBOLOS E ATRIBUIÇÕES *CREATIVE COMMONS*

Creative Commons | Atribuição | Uso não-comercial | Compartilhamento pela mesma licença | Vedada a criação de obras derivadas

| | | |
|---|---|---|
|  | **Atribuição By** | É a mais permissiva das licenças Creative Commons, permite a remixagem, distribuição e adaptação desde que atribuídos os devidos créditos ao criador original. |
|  | **Compartilhamento pela mesma licença** | Permite a remixagem e adaptação desde que atribuídos os devidos créditos ao criador original e que a distribuição seja sob a mesma licença. |
|  | **Atribuição sem obras derivadas** | Permite a distribuição, sem remixagem e adaptação, desde que atribuídos os devidos créditos ao criador mesmo que para fins comerciais e não comercias. |
|  | **Atribuição sem uso comercial** | Permite a distribuição, remixagem e adaptação, desde que atribuídos os devidos créditos ao criador exceto para fins comerciais. |
|  | **Atribuição compartilhada e sem uso comercial** | Permite a distribuição, remixagem e adaptação, desde que atribuídos os devidos créditos ao criador, exceto para fins comerciais e desde que sob a mesma licença. |
|  | **Atribuição não comercial sem obras derivadas** | É a mais restritiva dentre as licenças Creative Commons. Permite apenas a redistribuição, desde que atribuídos os devidos créditos ao criador, exceto para fins comerciais. |

FONTE: Adaptado de Creative Commons (2015).

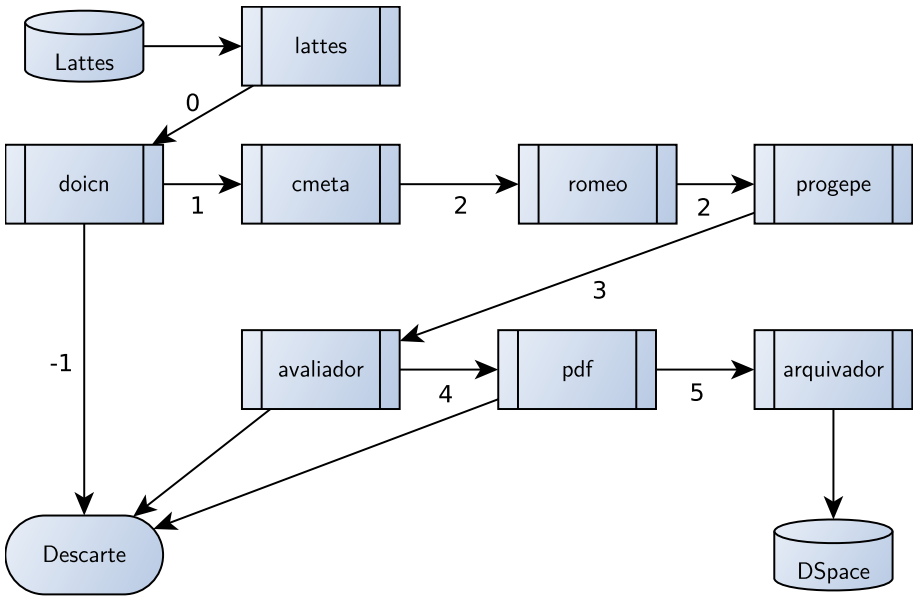
No capítulo seguinte é apresentada a metodologia que foi implementada para coletar publicações científicas e trazê-las para os repositórios digitais institucionais de acordo com a Iniciativa de Acesso Aberto.

# 8 COLETA E CLASSIFICAÇÃO DE ARTIGOS CIENTÍFICOS *OPEN ACCESS* EM RDIS

O objetivo desse trabalho foi o desenvolvimento de um conjunto de componentes, para coleta e classificação de metadados de artigos científicos produzidos sob uma licença de acesso aberto no repositório digital da Universidade Federal do Paraná. Os componentes foram executados de forma encadeada, e seus resultados armazenados em um banco de dados relacional.

Cada componente, criado com a linguagem Python, identifica e atualiza os DOIs para processamento de acordo com um estágio específico (0, −1, 1, 2, 3, 4 e 5), conforme a Figura 9. Os passos executados por cada componente são detalhados a seguir.

FIGURA 9 – FLUXO DE DADOS PARA SELEÇÃO DE ARTIGOS




FONTE: A autora (2016).

Como primeiro passo para implementação do estudo, a partir dos currículos extraídos da Plataforma Lattes, foram obtidos os DOIs e metadados de artigos publicados, constantes na *tag ARTIGO-PUBLICADO* (Figura 10). Os dados selecionados foram armazenados no banco de dados relacional, utilizando a linguagem de programação Python, com estágio 0.

FIGURA 10 – EXEMPLO DE EXTRAÇÃO DE DADOS XML DE UM ARTIGO PUBLICADO

```
<ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="26" ORDEM-IMPORTANCIA="">
 <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Uso de XML para
Interoperabilidade entre Bases Heterogêneas" ANO-DO-ARTIGO="2003" PAIS-DE-PUBLICACAO="Brasil"
IDIOMA="Português" MEIO-DE-DIVULGACAO="MEIO_DIGITAL" HOME-PAGE-
DO-TRABALHO="www.presidentekennedy.br" FLAG-RELEVANCIA="NAO" DOI="" TITULO-DO-ARTIGO-INGLES=""
FLAG-DIVULGACAO-CIENTIFICA="NAO"/>
 <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="RESI. Revista Eletrônica de Sistemas
de Informação" ISSN="16773071" VOLUME="II" FASCICULO="I" SERIE="2" PAGINA-INICIAL="1" PAGINA-
FINAL="12" LOCAL-DE-PUBLICACAO="brasil"/>
 <AUTORES NOME-COMPLETO-DO-AUTOR="Juliana Pasqual" NOME-PARA-CITACAO="Pasqual J." ORDEM-
DE-AUTORIA="2" NRO-ID-CNPQ=""/>
 <AUTORES NOME-COMPLETO-DO-AUTOR="Marcos Sfair Sunye" NOME-PARA-CITACAO="SUNYE, Marcos
Sfair;Sunye, Marcos S." ORDEM-DE-AUTORIA="1"/>
 <PALAVRAS-CHAVE PALAVRA-CHAVE-1="BANCO DE DADOS" PALAVRA-CHAVE-2="ENTIDADE RELACIONAMENTO"
PALAVRA-CHAVE-3="INTEGRACAO" PALAVRA-CHAVE-4="XML" PALAVRA-CHAVE-5="" PALAVRA-CHAVE-6=""/>
 <AREAS-DO-CONHECIMENTO>
  <AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA"
NOME-DA-AREA-DO-CONHECIMENTO="Ciência da Computação" NOME-DA-SUB-AREA-
DO-CONHECIMENTO="Metodologia e Técnicas da Computação" NOME-DA-ESPECIALIDADE=""/>
  <AREA-DO-CONHECIMENTO-2 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA"
NOME-DA-AREA-DO-CONHECIMENTO="Ciência da Computação" NOME-DA-SUB-AREA-
DO-CONHECIMENTO="Metodologia e Técnicas da Computação" NOME-DA-ESPECIALIDADE="Banco de Dados"/>
 </AREAS-DO-CONHECIMENTO>
 <SETORES-DE-ATIVIDADE SETOR-DE-ATIVIDADE-1="Informática" SETOR-DE-ATIVIDADE-2="" SETOR-
DE-ATIVIDADE-3=""/>
 <INFORMACOES-ADICIONAIS DESCRICAO-INFORMACOES-ADICIONAIS="" DESCRICAO-INFORMACOES-ADICIONAIS-
INGLES=""/>
</ARTIGO-PUBLICADO>
```

FONTE: A autora (2016).

Na sequência, os metadados dos DOIs foram requisitados junto ao resolvedor *dx.doi.org*, conforme exposto na Figura 11. Mediante validação dos DOIs, os dados foram armazenados em outra tabela no banco de dados (estágio 1), e os artigos com DOI inválido foram sinalizados para descarte (estágio −1).

FIGURA 11 – METADADOS DOI NO FORMATO `CITEPROC`

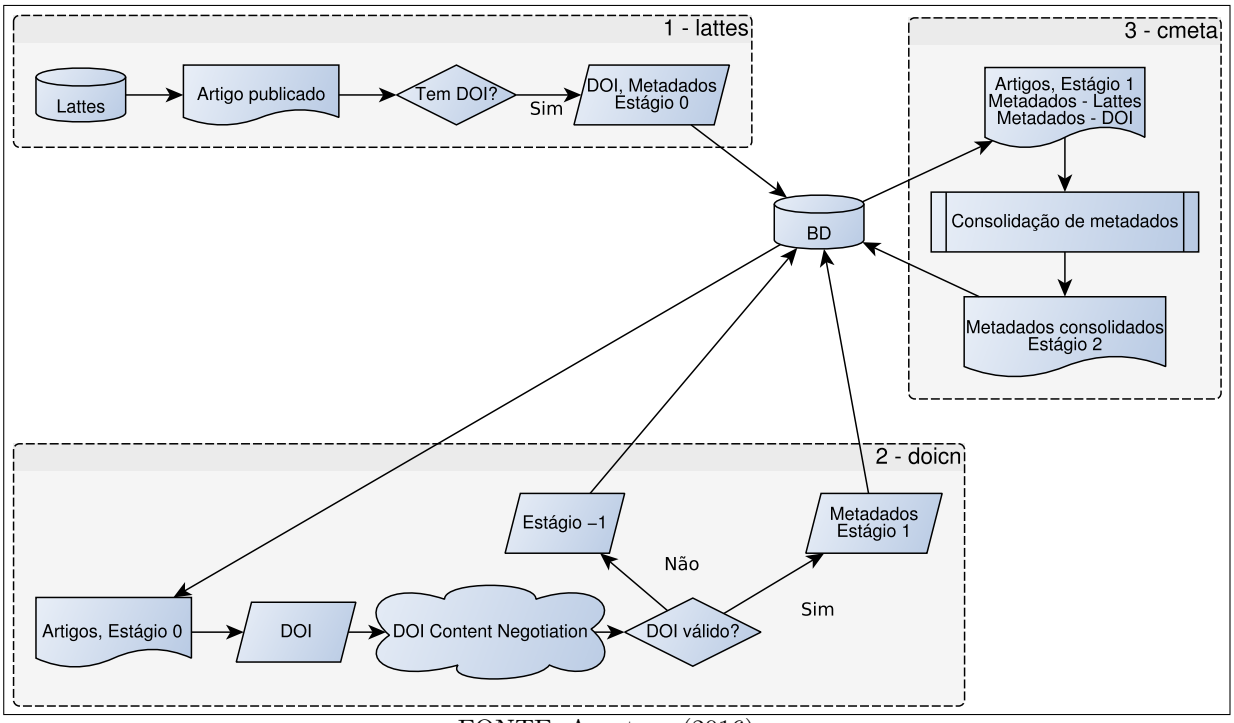
```
  "DOI": "10.1007/s00799-012-0080-5",
  "type": "journal-article",
  // [...]
  "ISSN": ["1432-5012", "1432-1300"],
  "title": "Extending OAI-PMH over structured P2P networks for digital preservation",
  "container-title": "Int J Digit Libr",
  "publisher": "Springer Science + Business Media",
  "volume": "12",
  "issue": "1",
  "issued": {
    "date-parts": [[2012, 2, 28]]
  },
  "author": [
    { "given": "Everton F. R.", "family": "Seára", "affiliation": [] },
    { "given": "Marcos S.", "family": "Sunye", "affiliation": [] },
    { "given": "Luis C. E.", "family": "Bona", "affiliation": [] },
    { "given": "Tiago", "family": "Vignatti", "affiliation": [] },
    { "given": "Andre L.", "family": "Vignatti", "affiliation": [] },
    { "given": "Anne", "family": "Doucet", "affiliation": [] }
  ]
}
```

FONTE: Adaptado de DOI (2013).

Na sequência, os conjuntos de metadados obtidos das bases Lattes e DOI (estágio 1) foram consolidados, considerando que os metadados da base DOI [1] são superiores aos da Lattes, e marcados para continuação (estágio 2). O motivo disto é que, por exemplo, o título do artigo com DOI *10.1023/A:1004939919783* está em branco na base de dados da CrossRef, porém é informado no currículo do docente na Plataforma Lattes.

A Figura 12 demonstra as etapas inicias para verificação se o artigo é de acesso aberto ou não.

FIGURA 12 – FLUXO DE ENTIDADES E RELACIONAMENTOS PARA SELEÇÃO DE ARTIGOS CIENTÍFICOS DE ACESSO ABERTO, ETAPAS 1–3




FONTE: A autora (2016).

Fez-se necessário verificar sob qual licença cada artigo no estágio 2 foi produzido através de uma consulta, utilizando o International Standard Serial Number (ISSN) do periódico envolvido, à base SHERPA/RoMEO, que é uma iniciativa para identificação de publicações científicas de acordo com o movimento de acesso aberto para obtenção de informações adicionais sobre sob qual licença o artigo foi publicado. Estas informações são disponibilizadas nas *tags preprints*, *postprints* e *pdfversion* do arquivo XML, conforme demonstrado na Figura 13.

De posse das informações obtidas foram armazenados em um banco de dados relacional os periódicos identificados como acesso aberto e, por uma referência cruzada foram classificadas os artigos pertencentes a estes periódicos.

[1] Os metadados retornados são os efetivamente depositados no sistema CrossRef.

FIGURA 13 – CONSULTA AO SHERPA/ROMEO RETORNANDO RESULTADOS DO DOAJ

```
<!-- [...] -->
<romeoapi version="2.9.9">
  <!-- [...] -->
  <publishers>
    <publisher id="DOAJ">
      <name>Tecpar</name>
      <homeurl>http://www.scielo.br/scielo.php?pid=1516-8913&
amp;script=sci_serial</homeurl>
      <preprints>
        <prearchiving>unknown</prearchiving>
        <prerestrictions />
      </preprints>
      <postprints>
        <postarchiving>unknown</postarchiving>
        <postrestrictions />
      </postprints>
      <pdfversion>
        <pdfarchiving>unknown</pdfarchiving>
        <pdfrestrictions />
      </pdfversion>
      <!-- [...] -->
      <romeocolour>gray</romeocolour>
    </publisher>
  </publishers>
</romeoapi>
```

FONTE: Adaptado de SHERPA/RoMEO (2015b).

Ainda com os artigos no estágio 2 e, para garantir apenas a identificação da produção científica dos docentes da universidade, considerando a possível produção de um docente inativo por atuação em outra instituição, foi feita uma consulta à base funcional institucional para identificação dos docentes pertencentes à universidade.

Para identificação das possíveis combinações de abreviação de nomes de autores comumente constantes nas bases utilizadas, foi utilizado um algoritmo que manteve o último sobrenome e gerou abreviações para os nomes restantes. Por exemplo, para os nomes dos autores “MARCOS SFAIR SUNYE” e “MAURICIO SUNYE” foram geradas as respectivas combinações:

- “MARCOS SFAIR SUNYE”, “M SFAIR SUNYE”, “MARCOS S SUNYE”, “MARCOS SUNYE”, “M S SUNYE”, e “M SUNYE”;
- “MAURICIO SUNYE” e “M SUNYE”.

Para vinculação dos artigos científicos aos autores da instituição, o algoritmo compara o nome dos mesmos às abreviações constantes na base.

O algoritmo não gera as combinações “M S S” ou “M S”, logo artigos sinalizados desta forma, ou nos casos onde sejam encontradas duas ocorrências da mesma abreviação, estes são descartados.

Neste momento, todos os artigos na base foram atualizados para o estágio 3 e, após isso, foi feita uma avaliação destes com base nos seguintes critérios:

- os metadados do artigo incluem ao menos um ISSN;
- ao menos um autor foi vinculado à instituição e o artigo foi publicado durante sua permanência na mesma; e
- a versão publicada do artigo pode ser arquivada.

Os artigos passíveis de serem armazenados no RDI da instituição foram, então, atualizados e colocados no estágio 4.

De posse disso, foram efetivamente coletados os arquivos Portable Document Format (PDF) dos artigos selecionados. Foram criados "extratores" para varrer o código HTML da *url* específica onde o objeto digital está armazenado, a fim de localizar o endereço para obtenção do objeto digital pretendido. Nesse momento foi identificado que, em alguns periódicos, é possível, a partir do próprio DOI, acessar o artigo em PDF. Para tratar dos DOIs nesta situação, foi necessário desenvolver um extrator específico.

Também foram identificados periódicos que não permitem a coleta dos artigos, seja por bloqueio de métodos automatizados (robôs) ou outros motivos, o que também foi resolvido mediante a criação de um extrator especifico.

Os artigos cujos PDFs são livres para coleta foram armazenados em um diretório específico e marcados no banco de dados como estágio 5.

De posse dos metadados e PDFs selecionados, foi montado uma estrutura de diretórios no formato *Simple Archive Format*[2] para importação pelo software DSpace.

[2] Disponível em: https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format

FIGURA 14 – EXEMPLO DE ARQUIVO DE METADADOS DE ACORDO COM O *SIMPLE ARCHIVE FORMAT*

```
<dublin_core>
  <dcvalue element="identifier" qualifier="other">10.1073/PNAS.0508170103</dcvalue>
  <dcvalue element="title">Linoleic acid hydroperoxide reacts with hypochlorous acid, generating peroxyl radical intermediates and singlet molecular oxygen</dcvalue>
  <dcvalue element="relation" qualifier="ispartof">Proceedings of the National Academy of Sciences, v. 103, n. 2</dcvalue>
  <dcvalue element="publisher">Proceedings of the National Academy of Sciences</dcvalue>
  <dcvalue element="date" qualifier="issued">2005-12-30</dcvalue>
  <dcvalue element="identifier" qualifier="issn">0027-8424</dcvalue>
  <dcvalue element="identifier" qualifier="issn">1091-6490</dcvalue>
  <dcvalue element="contributor" qualifier="author">D. Rettori</dcvalue>
  <dcvalue element="contributor" qualifier="author">G. R. Martinez</dcvalue>
  <dcvalue element="contributor" qualifier="author">M. H. G. Medeiros</dcvalue>
  <dcvalue element="contributor" qualifier="author">O. Augusto</dcvalue>
  <dcvalue element="contributor" qualifier="author">P. Di Mascio</dcvalue>
  <dcvalue element="contributor" qualifier="author">S. Miyamoto</dcvalue>
</dublin_core>
```

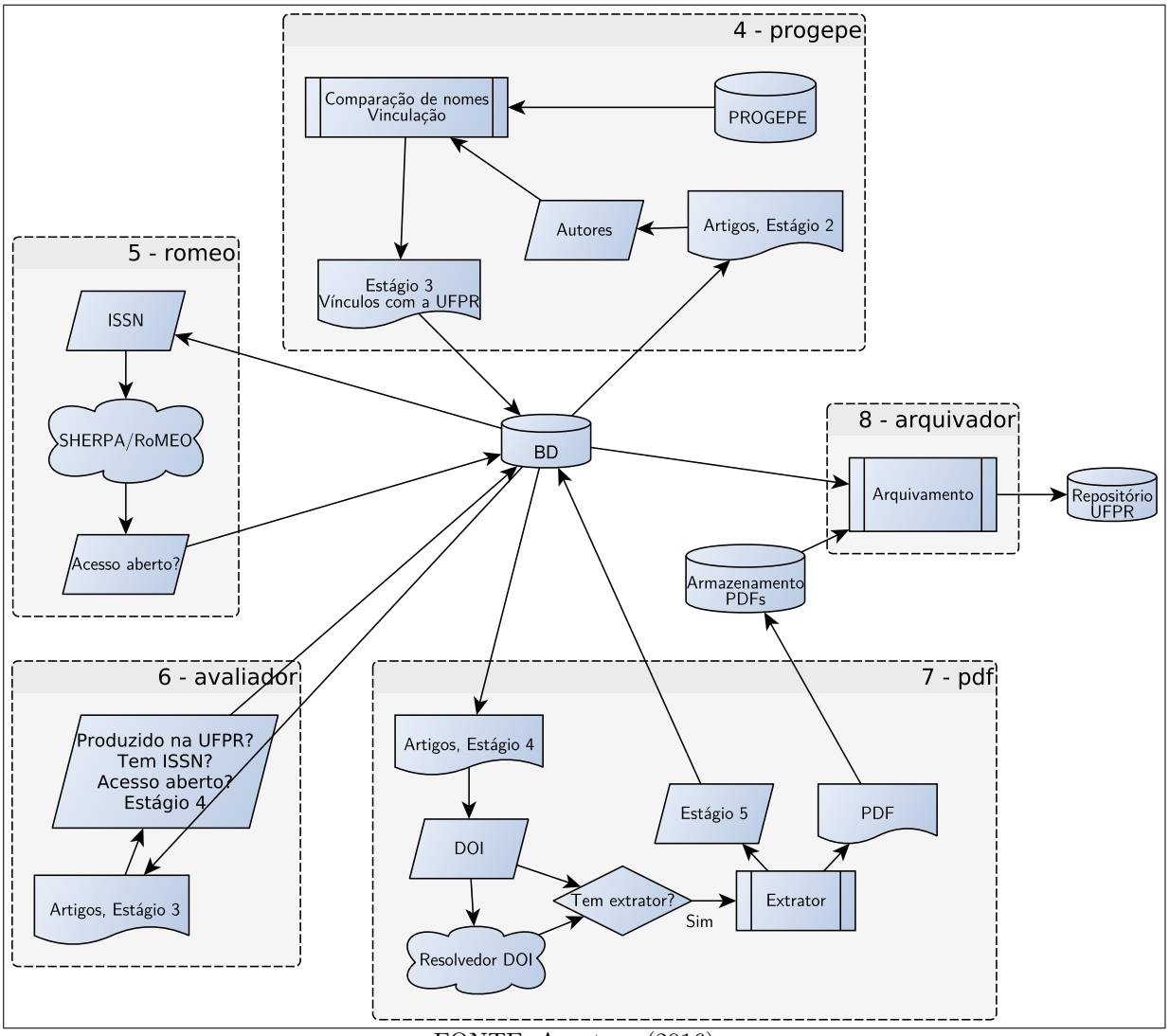
FONTE: A autora (2016).

Finalmente foi realizada a execução do API mencionado no parágrafo anterior como o comando abaixo para a carga dos PDFs dos artigos selecionados e seus respectivos metadados no RDI da instituição.

- dspace import -e usuario@inf.ufpr.br -a -c 1884/39 -s arquivo -m arquivo/map -R

A Figura 15 demonstra uniformemente todas as etapas referentes à identificação funcional e de periódicos em acesso aberto, avaliação dos dados coletados, coleta de PDFs e geração de arquivo para importação em RDIs.

FIGURA 15 – FLUXO DE ENTIDADES E RELACIONAMENTOS PARA SELEÇÃO DE ARTIGOS CIENTÍFICOS DE ACESSO ABERTO, ETAPAS 4–8



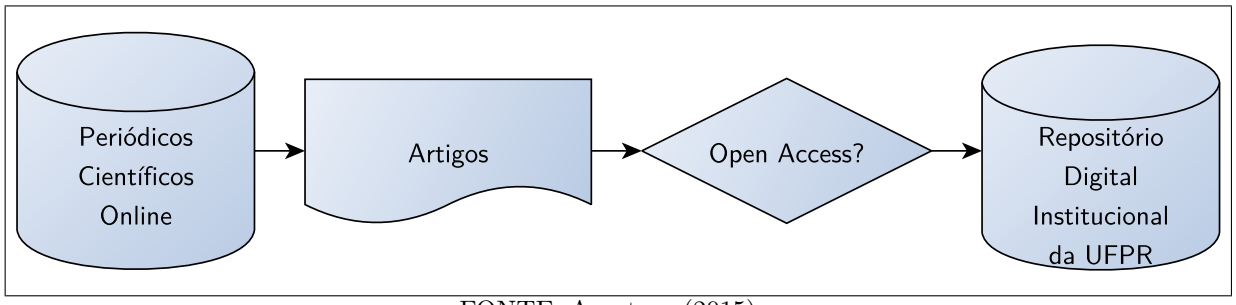

FONTE: A autora (2016).

## 8.1 ESTUDO DE CASO E ANÁLISE

Das pesquisas originam-se as publicações científicas que, "objetivam divulgar a pesquisa científica para a comunidade, de forma que permita que outros possam utilizá-la e avaliá-la sob outras visões" (BROFMAN, 2012).

A Figura 16 abaixo, mostra o processo de seleção dos artigos científicos que foram classificados para importação no RDI/UFPR:

FIGURA 16 – SELEÇÃO DE ARTIGOS CIENTÍFICOS PARA IMPORTAÇÃO




FONTE: A autora (2015).

A finalidade disso foi agregar os dados para posterior exportação dos metadados de acordo com o formato *Dublin Core* que é o esquema de metadados adotado pelo repositório digital objeto deste estudo.

A partir de uma extração da Plataforma Lattes datada de agosto de 2015, foram obtidos 1295 currículos pertencentes a docentes dos programas de pós-graduação da Universidade Federal do Paraná, o que resultou em 34.441 ocorrências/menções à artigos científicos contidos na *tag* ARTIGO-PUBLICADO. Destas menções, foram identificadas 8.969 que possuíam informação no atributo *DOI*, sendo destes 6.891 únicos e, que após submissão ao resolvedor, resultaram em 6.777 DOIs válidos.

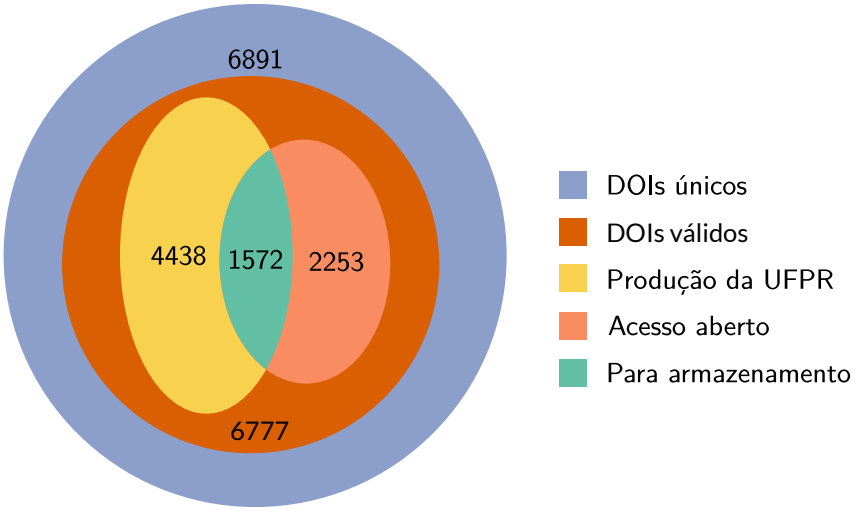
Os artigos com DOI válido referenciam 36.463 autores, dos quais 8.907 foram identificados como pertencentes ao quadro de docentes da universidade.

Os artigos fazem referência, ainda, a 2.783 ISSNs, dos quais 545 (19,6 %) pertencem a periódicos que permitem o arquivamento da versão publicada dos artigos, 2.029 (72,9 %) correspondem a periódicos que não permitem o arquivamento e 209 (7,5 %) não são listados no SHERPA/RoMEO.

A avaliação identificou que, do total de 6.777 artigos com DOI válido, 2.253 permitiam o arquivamento em outras bases de dados, como RDIs. Ainda desses 6.777, 4.438 artigos foram produzidos pelos então docentes da UFPR, sendo que destes, 1.572 artigos podiam ser replicados no repositório digital da referida instituição.

Este estudo identificou que, na Universidade Federal do Paraná, 35,4 % da produção científica mencionada na Plataforma Lattes, que tem DOI informado, foi disponibilizada em acesso aberto, conforme demonstrado na Figura 17.

FIGURA 17 – ARTIGOS IDENTIFICADOS PARA CARGA NO RDI/UFPR




FONTE: A autora (2016).

Uma vez identificado que um artigo científico foi disponibilizado sob uma licença de acesso aberto, foram coletados e armazenados em um banco de dados SQLite os metadados necessários para identificação do mesmo em um repositório digital, acrescidos de um metadado específico que indica qual a forma de licença atrelada ao mesmo.

A carga automatizada no repositório digital foi feita da seguinte forma: (1) criação de uma comunidade denominada "Produção Científica"; (2) criação de uma coleção chamada "Artigos" e; (3) importação dos dados e metadados, conforme Figura 18.

FIGURA 18 – COLEÇÃO DE ARTIGOS NO RDI/UFPR


FONTE: A autora (2016).

Os códigos fontes dos componentes [3] foram desenvolvidos de acordo com a filosofia de acesso aberto para compartilhamento e difusão do conhecimento científico produzido.

No próximo capítulo são demonstrados os resultados alcançados, as dificuldades encontradas e os meios utilizados para resolvê-las bem como, foram elencados, ainda, alguns tópicos para trabalhos futuros que possam dar continuidade a este trabalho.

[3] Disponível em: git@gitlab.c3sl.ufpr.br:elisabete/cmpc.git

# 9 CONCLUSÃO

A adoção de Repositórios Digitais Institucionais promove a disseminação de conteúdo técnico-científico produzido por uma instituição e enriquece culturalmente os que dela se beneficiam. A agregação da produção científica em um único local facilita o acesso a uma quantidade maior de informação e por consequência fomenta a transferência de conhecimento.

Além disso, a agregação da produção científica em um único ambiente digital institucional permite à instituição participante gerar indicadores internos de produção científica e tecnológica, realizar estudos através da aplicação de ferramentas de mineração de dados, e apoiar a implementação de políticas de gestão.

Um dos indicadores utilizados para classificação de bibliotecas ou repositórios digitais institucionais, por rankings universitários, é o volume de produção científica constante nestes ambientes.

Embora os autores de produções científicas sejam orientados a disponibilizarem seus trabalhos em um formato eletrônico normalizado em pelo menos um repositório de acesso aberto, de acordo com a *Berlin Declaration on Open Access* (2003), ainda é nítida a falta de sensibilização por parte dos mesmos em relação a este tópico.

Este trabalho também demonstrou a necessidade da sensibilização do corpo docente quanto a importância de possuir e manter atualizado um currículo na Plataforma Lattes bem como, informar corretamente os DOIs associados às suas publicações, por ser este um identificador único permanente do artigo para sua recuperação no ambiente *web*.

A partir deste trabalho a instituição poderá planejar os custos necessários para manutenção de seu ambiente digital, através da obtenção do volume de produção científica a ser armazenado em seu RDI.

## 9.1 DIFICULDADES ENCONTRADAS

- **Artigos sem DOI** - a certificação por DOI foi inserida na Plataforma Lattes apenas a partir de 2007 e também não é uma informação obrigatória, logo foram invalidados da base de dados obtida cerca de 25000 artigos científicos por não possuírem a informação desejada.
- **Artigos com DOI inválido** - Como a certificação por DOI foi inserida na Plataforma Lattes apenas a partir de 2007, o campo DOI não possuía validação quanto ao conteúdo informado, logo foram registrados 114 DOIs (1.7 %) com formato inválido.
- **Identificação de Homônimos** - Embora, já existam iniciativas para tratamento envolvendo homônimos e abreviações como o Open Researcher and Contributor

ID (ORCID) a adesão a estes serviços, embora gratuita, ainda é incipiente por parte dos autores de produção científica.

- **Artigos de outras instituições** - O código do profissional junto a Plataforma Lattes é único e para a obtenção do arquivo XML a busca é realizada por CPF, logo é retornada toda a produção científica do profissional em todas as instituições onde o mesmo tenha tido vínculo profissional.

  Outra dificuldade é a falta de sensibilização de autores em cadastrar e manter atualizado junto ao ORCID seus dados funcionais, bem como adotar o hábito de adicionar um metadado, que indique seu código ORCID, a fim de permitir a correta identificação quanto a sua afiliação no momento da publicação da produção científica.

  Foi necessário criar uma ferramenta para selecionar e tratar somente os artigos científicos pertencentes aos docentes da UFPR.

- **Padronização** - Há falta de padronização para busca e obtenção dos PDFs envolvidos. Para tanto, foi preciso criar extratores específicos por periódicos, conforme exemplificado na Figura 19.

- **Artigos pagos** - Alguns periódicos estão cadastrados no SHERPA/RoMEO como "acesso aberto total" embora exijam pagamento para acesso às suas publicações. Por este motivo foi preciso criar um extrator específico para bloqueio dos prefixos destes periódicos, conforme Figura 19.

- **Bloqueio de robôs** - Existem periódicos que são de acesso aberto, porém não permitem acesso às suas publicações por meio de robôs. Por este motivo foi preciso criar um extrator específico para bloqueio dos prefixos destes periódicos. A Figura 19 exemplifica a situação.

FIGURA 19 – EXEMPLOS DE EXTRATORES: BLOQUEIO, PDF E GOOGLE SCHOLAR

```
class Bloqueios(metaclass=Extratores):
    prefixos_bloqueados = {
        '10.1038',  # Nature: keeps disconnecting us
        '10.1039',  # RSC: not a PDF (also, not open access)
        '10.1063',  # AIP: relative URLs don't work when testing
        '10.1103',  # APS: ToS
        '10.1107',  # IUC: Password-protected
        '10.1186',  # BioMed Central: disconnecting
    }

    @classmethod
    def extrair(cls, doi, resposta, html, http):
        if DOI.tem_prefixo(doi, cls.prefixos_bloqueados):
            raise StopIteration


class ExtratorPDF(metaclass=Extratores):
    @classmethod
    def extrair(cls, doi, resposta, html, http):
        if resposta.headers['Content-Type'] in ('application/pdf', 'application/x-pdf'):
            return cls(doi, pdf=resposta.content)


class ExtratorGoogleScholar(metaclass=Extratores):
    @classmethod
    def extrair(cls, doi, resposta, html, http):
        citation_pdf_url = html.xpath('string(//meta[@name="citation_pdf_url"]/@content)')
        if citation_pdf_url:
            return cls(doi, url_pdf=cls.URL(html, citation_pdf_url))
```

FONTE: A autora (2016).

Este trabalho limitou-se em desenvolver um conjunto de componentes, para coleta e classificação de metadados de artigos científicos produzidos em acesso aberto no RDI/UFPR, porém é importante salientar, conforme já afirmado nos capítulos 1 e 5, a relevância de se agregar em um mesmo ambiente toda a produção científica de uma instituição. Logo, foram elencados os seguintes itens de produção científica a serem classificados e agregados no RDI em desenvolvimento futuro:

- classificação e seleção de metadados de eventos científicos;
- classificação e seleção de metadados de atividades de extensão universitária.

# REFERÊNCIAS

ALVES, Maria das Dores Rosa; SOUZA, Marcia Izabel Fugisawa. Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 4, n. 2, p. 20–38, 2007. ISSN: 1678-765X. Disponível em: <http://www.sbu.unicamp.br/seer/ojs/index.php/rbci/article/view/358>. Acesso em: 8 out. 2015.

APS NEWS. **APS forms cooperative agreement with LANL's xxx e-print archive**. [S.l.: s.n.], 1998. Disponível em: <http://web.archive.org/web/20020113112838/http://positron.aps.org/apsnews/0398/039805.html>. Acesso em: 29 mar. 2016.

*ARXIV*. Disponível em: <http://arxiv.org/>. Acesso em: 11 jul. 2016.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 10520**: informação e documentação – citações em documentos – apresentação. Rio de Janeiro, 2002.

BACA, Murtha. (Ed.). **Introducción a los metadatos**: vías a la información digital. Tradução de Marisol Jacas-Santoll. [S.l.]: Getty, 1998. ISBN: 0-89236-535-8.

BARRETO, Aldo de Albuquerque. A transferência de informação, o desenvolvimento tecnológico e a produção de conhecimento. **Questões em Rede**, v. 1, p. 1, 2012.

*BERLIN DECLARATION ON OPEN ACCESS*. 2003. Disponível em: <http://openaccess.mpg.de/Berlin-Declaration>. Acesso em: 24 mar. 2016.

BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. **Diário Oficial da União**, Brasília, p. 3, 20 fev. 1998. Seção 1. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9610.htm>.

BROFMAN, Paulo Roberto. A importância das publicações científicas. **Cogitare Enfermagem**, v. 17, n. 3, p. 419, 2012.

*BUDAPEST OPEN ACCESS INITIATIVE*. 2002. Disponível em: <http://www.budapestopenaccessinitiative.org/read>. Acesso em: 29 mar. 2016.

CAMARGO, Liriane Soares de Araújo de; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Arquitetura da informação para repositórios digitais. In: SAYÃO, Luis et al. **Implantação e gestão de repositórios institucionais**: políticas, memória, livre acesso e preservação. Salvador: EDUFBA, 2009. cap. 3, p. 57–82.

CASTRO, Cristiane Yanase Hirabara de et al. Repositórios institucionais confiáveis: repositório institucional como ferramenta para a preservação digital. In: SAYÃO, Luis et al. **Implantação e gestão de repositórios institucionais**: políticas, memória, livre acesso e preservação. Salvador: EDUFBA, 2009. cap. 13, p. 283–304.

CITY RESEARCH ONLINE. **FAQ & Contacts**: How does the system work. [S.l.: s.n.]. Disponível em: <http://openaccess.city.ac.uk/>. Acesso em: 11 jun. 2015.

CORK OPEN RESEARCH ARCHIVE. **How to submit your research**. [S.l.: s.n.]. Disponível em: <http://cora.ucc.ie/>. Acesso em: 11 jun. 2015.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES. **Handle.net**. Disponível em: <http://www.handle.net/HNRj/HNR-Policies.pdf>. Acesso em: 18 jun. 2015.

CREATIVE COMMONS. **Sobre as licenças**. Disponível em: <https://creativecommons.org/licenses/>. Acesso em: 13 mar. 2015.

DIADORIM. Disponível em: <http://diadorim.ibict.br/>. Acesso em: 25 nov. 2014.

DIGITALA VETENSKAPLIGA ARKIVET. **About DiVA portal**. [S.l.: s.n.]. Disponível em: <http://www.diva-portal.org/smash/aboutdiva.jsf>. Acesso em: 11 jun. 2015.

DIRECTORY OF OPEN ACCESS JOURNALS. Disponível em: <https://doaj.org/>. Acesso em: 25 nov. 2014.

DOI SYSTEM PROXY SERVER. **[Consulta ao DOI 10.1007/s00799-012-0080-5 com negociação de conteúdo]**. 2013. Disponível em: <https://doi.org/10.1007/s00799-012-0080-5>. Acesso em: 12 fev. 2016.

DUBLIN CORE METADATA INITIATIVE. **Dublin Core Metadata Element Set, Version 1.1**. [S.l.: s.n.], jun. 2012. Disponível em: <http://dublincore.org/documents/2012/06/14/dces/>. Acesso em: 11 nov. 2015.

DULCINEA. Disponível em: <http://www.accesoabierto.net/dulcinea>. Acesso em: 25 nov. 2014.

FERREIRA, Elisabete et al. Digital Object Identifier (DOI): O que é, para que serve, como se usa? **AtoZ: novas práticas em informação e conhecimento**, v. 4, n. 1, 2015.

*HAL*. Disponível em: <https://hal.archives-ouvertes.fr/>. Acesso em: 11 jul. 2016.

HARNAD, Stevan. **The research-impact cycle**. [S.l.: s.n.]. Disponível em: <http://opcit.eprints.org/feb19oa/harnad-cycle.ppt>. Acesso em: 15 abr. 2015.

INDECS. **Interoperability of Data in E-commerce Systems**: Overview. Dez. 1999. Disponível em: <https://web.archive.org/web/20000817054843/http://www.indecs.org/overview/overview.htm>. Acesso em: 11 nov. 2015.

INTERNATIONAL ASSOCIATION OF SCIENTIFIC, TECHNICAL AND MEDICAL PUBLISHERS. **Open Access Licensing**. Disponível em: <http://www.stm-assoc.org/copyright-legal-affairs/licensing/open-access-licensing/>. Acesso em: 12 nov. 2015.

INTERNATIONAL DOI FOUNDATION. **DOI Handbook**. Disponível em: <http://www.doi.org/doi_handbook/1_Introduction.html#1.2>. Acesso em: 18 jun. 2015.

JUNGMANN, Diana de Mello; BONETTI, Esther Aquemi. **A caminho da inovação: Proteção e Negócios com Bens de Propriedade Intelectual**: Guia para o Empresário. Brasília: IEL, 2010.

LANGIANO, Beatriz do Carmo. **Um Mecanismo para Automatizar a Criação dos Metadados das Imagens de Bibliotecas Digitais e Prover Buscas por Conteúdo**. 2005. 58 f. Dissertação (Mestrado em Informática) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2005.

LEITE, Fernando César. **Como gerenciar e ampliar a visibilidade da informação científica brasileira**: repositórios institucionais de acesso aberto. Brasília: IBICT, 2009.

LIBRARY OF CONGRESS. **About EAD**. Jul. 2012. Disponível em: <https://www.loc.gov/ead/eadabout.html>. Acesso em: 11 nov. 2015.

_____. **MARC Standards**. Jul. 2015. Disponível em: <https://www.loc.gov/marc/>. Acesso em: 11 nov. 2015.

_____. **MODS: Uses and Features**. Dez. 2014. Disponível em: <https://www.loc.gov/standards/mods/mods-overview.html>. Acesso em: 11 nov. 2015.

LIBRE ACCÈS AUX RAPPORTS SCIENTIFIQUES ET TECHNIQUES. **Qu'est-ce que le Dublin Core?** [S.l.: s.n.]. Disponível em: <http://lara.inist.fr/lara>. Acesso em: 28 maio 2015.

LIMA, Marcia H. T. de Figueredo. Consequências do movimento pelo livre acesso—open access—e o direito à informação científica. In: SAYÃO, Luis et al. **Implantação e gestão de repositórios institucionais**: políticas, memória, livre acesso e preservação. Salvador: EDUFBA, 2009. cap. 9, p. 219–230.

LUCE, Richard E. E-prints Intersect the Digital Library: Inside the Los Alamos arXiv. **Issues in Science and Technology Librarianship**, 2001. DOI: 10.5062/F44B2Z95. Disponível em: <http://www.istl.org/01-winter/article3.html>. Acesso em: 22 mar. 2016.

MAHASSEN, Naadin. **A quantitative approach to world university rankings**. Center for World University Rankings. 2014. Disponível em: <http://cwur.org/methodology/preprint.pdf>. Acesso em: 25 nov. 2014.

MEADOWS, Arthur Jack. **A Comunicação Científica**. Brasília: Briquet de Lemos, 1999.

MURAKAMI, Tiago Rodrigo Marçal. **Contato sobre o repositório institucional da USP**. [S.l.: s.n.], 24 mar. 2015. [mensagem pessoal]. Mensagem recebida por ferreira.bete@gmail.com.

NATIONAL ACADEMIC RESEARCH AND COLLABORATIONS INFORMATION SYSTEM. Disponível em: <http://www.narcis.nl/?Language=en>. Acesso em: 18 jun. 2015.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **A Framework of Guidance for Building Good Digital Collections**. 2007. Disponível em: <http://www.niso.org/publications/rp/framework3.pdf>. Acesso em: 17 jun. 2015.

_____. **Understanding Metadata**. Bethesda, MD, EUA: NISO, 2004. ISBN: 1-880124-62-9.

NATIONAL UNIVERSITY OF IRELAND GALWAY REPOSITORY. **Deposit Guide**. Disponível em: <http://aran.library.nuigalway.ie/xmlui/depositguide.html>. Acesso em: 11 jun. 2015.

OLIVEIRA, Renan Rodrigues de. **Recuperação Contextualizada de Documentos Integrados pelo Protocolo OAI-PMH**. 2010. 74 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2010.

OPEN ARCHIVES INITIATIVE. **About OAI**. Disponível em: <http://www.openarchives.org/OAI/OAI-organization.php>. Acesso em: 30 mar. 2016.

PAVÃO, Caterina Groposo; COSTA, Janise Silva Borges da et al. Motivações e desafios para a criação do repositório digital da Universidade Federal do Rio Grande do Sul, 2013.

PAVÃO, Caterina Groposo; SOUSA, Rodrigo Silva Caxias de; CAREGNATO, Sônia Elisa. Publicização da literatura científica através de repositórios institucionais. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 23., 2009, Bonito, MS. **Anais...** São Paulo: FEBAB, 2009. Não paginado.

PLATAFORMA LATTES. **Comparativo por Geografia, Instituição de Vínculo e Área de Atuação**. 2015. Disponível em: <http://estatico.cnpq.br/painelLattes/comparacao/>. Acesso em: 17 mar. 2015.

_____. **Dados e Estatísticas**. 2016. Disponível em: <http://memoria.cnpq.br/web/portal-lattes/dados-e-estatisticas>. Acesso em: 11 maio 2016.

_____. **Sobre a Plataforma**. 2012. Disponível em: <http://www.cnpq.br/web/portal-lattes/sobre-a-plataforma>. Acesso em: 17 mar. 2015.

PLOS ADVOCACY. **Coalition Letter on STM Model Licenses**. Ago. 2014. Disponível em: <https://dx.doi.org/10.6084/m9.figshare.1130855>. Acesso em: 12 nov. 2015.

RAMOS, Renan Carvalho; ANDRETTA, Pedro Ivo Silveira; SILVA, Eduardo Graziosi. Considerações Acerca do Processo de Alimentação de Repositórios Através da Importação de Registros de Bases de Dados Internacionais. **RDBCI**, v. 10, p. 1, 2012.

ROCHA, Hilton Ricardo. Software & Direito—Dos contratos: Licença de uso e serviços. **Âmbito Jurídico**, v. 42, jun. 2007. Disponível em: <http://www.ambito-juridico.com.br/site/index.php?n_link=revista_artigos_leitura&artigo_id=1914>. Acesso em: 4 nov. 2015.

SANTOS, Raimundo Nonato Macedo dos; KOBASHI, Nair Yumiko. Aspectos Metodológicos da Produção de Indicadores em Ciência e Tecnologia. In: VI ENCONTRO NACIONAL DE ENSINO E PESQUISA EM INFORMAÇÃO, Salvador, BA. **Anais...** [S.l.]: Universidade Federal da Bahia, 2005. Disponível em: <http://www.cinform-anteriores.ufba.br/vi_anais/docs/RaimundoNonatoSantos.pdf>. Acesso em: 5 abr. 2016.

SAYÃO, Luis Fernando. Interoperabilidade das bibliotecas digitais: O papel dos sistemas de identificadores persistentes—URN, PURL, DOI, Handle System, CrossRef e OpenURL. **TransInformação**, v. 19, n. 1, p. 65–82, 2007a. doi:10.1590/s0103-37862007000100006.

_____. **Metadados para preservação digital**: Aplicação do modelo OAIS. 2007b. Disponível em: <http://www.documentoseletronicos.arquivonacional.gov.br/Media/publicacoes/ctdemetadadospreservacaodigitalsayao.pdf>. Acesso em: 11 nov. 2015.

SETENARESKI, Ligia Eliana. **Repositórios Digitais Abertos**: Um Movimento do Livre Acesso Alternativo à Estrutura Oligopolizada das Editoras Científicas. 2013. 113 f. Dissertação (Mestrado em Políticas Públicas) – Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, Curitiba, 2013.

SHERPA/ROMEO. Disponível em: <http://www.sherpa.ac.uk/romeo/>. Acesso em: 25 nov. 2014.

_____. **[Consulta à API pelo ISSN 1050-2947]**. 2015a. Disponível em: <http://www.sherpa.ac.uk/romeo/api29.php?issn=1050-2947>. Acesso em: 12 fev. 2016.

_____. **[Consulta à API pelo ISSN 1516-8913]**. 2015b. Disponível em: <http://www.sherpa.ac.uk/romeo/api29.php?issn=1516-8913>. Acesso em: 12 fev. 2016.

VALMORBIDA, Willian; WOLF, Alexandre Stürmer; MONTEIRO, Ana Paula Lisboa. Análise e implementação de um sistema integrado de busca a partir dos protocolos: OAI-PMH, Z39.50, SRW e SRU. **Novas Tecnologias na Educação**, v. 9, n. 2, 2011.

WADHAM, Rachel. The open archives metadata harvesting protocol. **Library Mosaics**, v. 13, n. 4, p. 20, 2002.

WEBOMETRICS. **Ranking Web of Repositories**. [S.l.: s.n.]. Disponível em: <http://repositories.webometrics.info/en>. Acesso em: 25 nov. 2014.

_____. **Ranking Web of Repositories**: Brazil. [S.l.: s.n.]. Disponível em: <http://repositories.webometrics.info/en/Latin_America/Brazil>. Acesso em: 12 jun. 2015.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. ***Berne Convention for the Protection of Literary and Artistic Works***. Revisão de Paris. 1971. Disponível em: <http://www.wipo.int/treaties/en/ip/berne/>. Acesso em: 5 abr. 2016.

_____. **Copyright**. Disponível em: <http://www.wipo.int/copyright/en/>. Acesso em: 13 mar. 2015.

YAMAOKA, Eloi; GAUTHIER, Fernando. Objetos digitais: Em busca da precisão conceitual. **Informação & Informação**, v. 18, n. 2, 2013.