Catálogo de dados dos trabalhos científicos de gestão ambiental e saúde da Escola Nacional de Saúde Pública Sérgio Arouca (ENSP/FIOCRUZ): Proposta

Cristiane Rozeno Parangaba
Especialista em Informação Científica e Tecnológica em Saúde no ICICT- Fiocruz
Tecnologista em Saúde Pública da Fundação Oswaldo Cruz
E-mail: parangaba@ensp.fiocruz.br

RESUMO:
Este trabalho tem como proposta a organização dos trabalhos científicos da linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública, da Escola Nacional de Saúde Pública (ENSP). Para tanto, objetiva-se a criação de um catálogo de conjuntos de dados brutos com informações descritivas sobre os dados como, por exemplo, seu conteúdo, abrangência temporal e geográfica, e qualidade, que auxiliará os usuários a análise preliminar sem a necessidade de adquiri-los. A partir de uma pesquisa de natureza exploratória das publicações dessa linha de pesquisa nos repositórios institucional e temático, e da relação semântica existentes entre eles, a meta é futuramente inserir os conjuntos de dados brutos de pesquisa identificados e classificados no catálogo, em um repositório de dados científicos que utilize um padrão de metadados.

Palavras-chave: Bases de dados científicos. Repositórios digitais. Repositórios institucionais. Catálogos.

ABSTRACT:
This work proposes the organization of scientific work of the research line Environmental Management and Health of the Public Health program, the National School of Public Health (ENSP). Therefore, the goal is to create a catalog of sets of raw data with descriptive information about the data, for example, its content, temporal and geographic scope, and quality, which will help users to preliminary analysis without the need to acquire them. From exploratory research publications this line of research in the institutional and thematic repositories, and existing semantic analysis between them, the goal is to eventually enter, the sets of raw data identified search and classified in the catalog, in a repository scientific data using a metadata standard.

Keywords: Scientific databases. Scientific data. Digital repositories. Institutional repositories. Catalogs.

1 INTRODUÇÃO

Na última década do século XX, a comunidade científica presenciou a derrubada de barreiras existentes do acesso à informação científica. Dentre elas, o

alto custo das assinaturas das revistas, dificuldade de acesso às informações e a distância geográfica entre as comunidades científicas. Com o avanço de Tecnologias da Comunicação e Informação (TIC) e a crise dos periódicos, essas barreiras foram derrubadas e surgiram novas possibilidades de acesso e disseminação das pesquisas.

O uso das tecnologias de informação para disseminação da pesquisa obteve um expressivo crescimento nas interações das comunidades científicas. Os pesquisadores espalhados geograficamente que estudam ou não o mesmo assunto, através da internet, passam a trabalhar em conjunto de forma a contribuir para a geração de outros conhecimentos e, consequentemente, para o aumento exponencial do volume de informações e o avanço da ciência. Corroborando com esses aspectos, Baptista e colaboradores afirmam que:

> o uso da comunicação eletrônica tem permitido, nas duas últimas décadas, que esses pesquisadores tanto realizem pesquisas em colaboração, quanto publiquem em coautoria, mesmo nos casos em que nunca tenham se encontrado pessoalmente (BAPTISTA *et al.*, 2007, p. 4) .

A crise dos periódicos foi causada pelo o aumento abusivo das assinaturas dos periódicos e que tornou inviável as bibliotecas renovarem suas revistas. Para algumas revistas, houve aumentos de 1 mil por cento entre 1989 e 2001 (CHALHUB *et al.*, 2012).

Kuramoto afirma que na crise dos periódicos “os pesquisadores de diversas partes do globo terrestre se reuniram e deram início a um grande movimento global em direção ao acesso aberto[1] à informação científica” (KURAMOTO, 2009, p. 7).

Após a reunião do *Open Society Institute* (OSI) em 2001, surge o primeiro documento oficial do movimento de acesso livre, *o Budapest Open Access Iniciative (BOAI)*[2] que define os princípios e as estratégias para a implantação e garantia de acesso livre à informação (CHALHUB *et al.*, 2012).

A *Budapest Open Acess Iniciative* recomendou duas estratégias complementares para que a literatura científica esteja disponível e acessível: a via Dourada, em periódicos científicos, disseminados sem restrições de acesso e uso, e a via Verde, em repositórios institucionais de acesso livre, através do auto arquivamento (LEITE, 2009). Outros documentos similares, como a Declaração de

[1] Segundo o *Open Acess* presente no Documento de Budapeste - acesso gratuito e sem barreiras aos resultados de pesquisas científicas via internet, sem distinção entre acesso livre e acesso aberto.
[2] De BUDAPEST OPEN ACCESS INITIATIVE, 2002.Disponível em:<http://www.budapestopenaccessinitiative.org/read> Acesso em: 07 out. 2015.

Bethesda[3] e a Declaração de Berlim[4], também são resultados de movimentos em favor do acesso livre ao conhecimento científico.

CHALHUB *et al*. (2012) sublinham que:

> o depósito dos resultados de pesquisa em repositórios de universidades ou institutos de pesquisa se faz necessário uma vez que a publicação em periódicos com avaliação por pares não é condição suficiente para que os resultados das pesquisas tenham seu impacto maximizado (CHALHUB *et al*.,2012, p. 161).

Segundo a Declaração de Berlim, os repositórios institucionais de acesso de acesso livre deverão constar resultados de pesquisas originais, dados de pesquisas não processados, metadados, fontes originais, representações digitais de materiais pictóricos, gráficos e material acadêmico multimídia (DECLARAÇÃO..., 2003).

No cenário Brasileiro, o acesso livre se deu através dos seguintes documentos: Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica[5], Declaração de Salvador sobre acesso aberto[6], Carta de São Paulo[7] e Declaração de Florianópolis[8].O Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)[9] além de ter lançado o Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica, tem comandado várias ações para implantar o acesso livre no Brasil. Entre elas estão: o desenvolvimento de projetos de Publicações Periódicas de Acesso Livre (PCAL) junto à Financiadora de Estudos e Projetos (FINEP), subscrevendo e reforçando toda argumentação em favor do acesso livre na assinatura da Declaração de Berlim, distribuição de tecnologias; como o software para construção e gestão de revistas científicas, Sistema Eletrônico de Editoração de Revistas (SEER) e a construção de repositórios institucionais e temáticos de acesso livre (KURAMOTO, 2008).

[3] Disponível em:<http://legacy.earlham.edu/~peters/fos/bethesda.htm> Acesso em: 07 out. 2015.
[4] Disponível em:<http://openaccess.mpg.de/Berlin-Declaration> Acesso em: 07 out. 2015.
[5] Disponível em: <https://bvc.cgu.gov.br/manifesto.jsp> Acesso em: 7 out. 2015.
[6] Disponível em: <http://blog.scielo.org/blog/2015/10/23/declaracao-de-salvador-sobre-acesso-aberto-completa-10-anos/>. Acesso em: 7 out. 2015.
[7] Disponível em:<http://www.ibict.br/Sala-de-Imprensa/noticias/2005/carta-de-sao-paulo-defende-o-acesso-livre-a/?searchterm=carta%20de%20sao%20paulo>. Acesso em: 7 out. 2015.
[8] Disponível em: <http://www.ibict.br/Sala-de-Imprensa/noticias/2006/declaracao-de-florianopolis/impressao>. Acesso em: 7 out. 2015.
[9] Disponivel em:<http://www.ibict.br/> Acesso em: 7 out. 2015.

1.1 O PAPEL DA FIOCRUZ NO MOVIMENTO DO ACESSO LIVRE

A Fundação Oswaldo Cruz (FIOCRUZ)[10] é composta por várias unidades técnicas científicas que possuem autonomias de atuações em diversas atividades. No que tange as iniciativas do tema movimento do acesso aberto[11], a primeira ocorre em abril de 2011, no Instituto de Comunicação e Informação Científica em Saúde (Icict/Fiocruz) que lançou o ARCA[12], o Repositório Institucional da Fiocruz.

A segunda iniciativa ocorreu em setembro de 2012 pela unidade Escola Nacional de Saúde Pública - ENSP lançou o seu repositório institucional de Produção Científica com a sua Política Institucional de Acesso Aberto ao Conhecimento[13], com o propósito de dar visibilidade à produção científica da escola em saúde pública.

A terceira iniciativa foi aprovada a Política Institucional de Acesso Aberto ao Conhecimento[14] para todas as unidades da FIOCRUZ em 2014, pensando em ampliar sua visibilidade científica e contribuir com o desenvolvimento da ciência e preservar sua produção,

A ENSP se dedica à formação profissional em saúde e ciência & tecnologia e atua, de forma protagonista, em pesquisa, desenvolvimento tecnológico, formulação de políticas públicas e prestação de serviços de referência em saúde. Além disso, produz informação e tem a responsabilidade de garantir acesso pleno ao conhecimento seja pelo sistema de biblioteca multimídias, ou por meio do Repositório em Saúde Pública, dentre outros (http://www.ensp.fiocruz.br/portal-ensp/).

O Seminário Internacional Acesso Livre ao Conhecimento (SEMINÁRIO..., 2011) foi o marco inicial de sua adesão ao Movimento Internacional de Acesso Aberto ao Conhecimento. Desde então, vem incentivando seus pesquisadores a abraçar de forma plena o depósito de suas publicações científicas, através do auto arquivamento, no repositório temático.

[10] Disponível em:<http://portal.fiocruz.br/pt-br/content/unidades-e-escritórios> Acesso em: 7 out. 2015.
[11] Segundo o *Open Acess* presente no Documento de Budapeste - acesso gratuito e sem barreiras aos resultados de pesquisas científicas via internet, sem distinção entre acesso livre e acesso aberto.
[12] Disponível em: https://portal.fiocruz.br/sites/portal.fiocruz.br/files/documentos/portaria_-_politica_de_acesso_aberto_ao_conhecimento_na_fiocruz.pdf Acesso em: 7 out. 2015.
[13] Disponível em:< http://www6.ensp.fiocruz.br/repositorio/node/368239> Acesso em: 7 out. 2015.
[14] Disponível em:<http://portal.fiocruz.br/pt-br/content/na-fiocruz> Acesso em: 7 out. 2015.

1.2 A IMPORTÂNCIA DOS REPOSITÓRIOS DE DADOS CIENTÍFICOS PARA A FIOCRUZ E PARA ENSP.

Há séculos, a produção dos dados científicos, que são as fontes primárias da pesquisa, se tornou importante para a pesquisa científica. Eles são produzidos e utilizados no contexto da pesquisa científica e estão evoluindo naturalmente em suas formas e volumes, crescendo em dimensão e complexidade, observam Rodrigues *et al.* (2010).

A Fiocruz em seu repositório institucional ARCA possui cerca de quase 10000[15] produções intelectuais, que são: artigos, capítulos de livros, dissertações, trabalhos de conclusão de cursos, relatórios, manuais e procedimentos técnicos e teses. De acordo com a natureza dessas produções, podem existir dados de pesquisas que serviram como base para os estudos e que podem estar armazenados em lugares desconhecidos ou mesmo de forma incorreta. Muitos poderão ser perdidos ou esquecidos, e daqui alguns anos as formas de armazenamento estarão ultrapassadas pela tecnologia.

Disponibilizar os dados de pesquisas a outros pesquisadores é importante em vários aspectos como: minimizar custos, permitir novos estudos utilizando esses dados e eliminar tempo gasto em projetos novos.

O movimento de acesso aberto pode contribuir também para quebrar as barreiras de acesso a esses dados e permitir que sejam compartilhados entre os pesquisadores para uso e reuso. Uma das formas para que esses dados sejam armazenados, preservados e acessados hoje e no futuro, é alocá-los em repositórios de dados.

Esse trabalho propõe a Fiocruz e inicialmente a ENSP organizar todos os dados brutos disponibilizados da linha de pesquisa Gestão Ambiental e Saúde para serem depositados em um repositório de dados científicos com o propósito de preservar e disponibilizar à comunidade científica e a sociedade.

2 JUSTIFICATIVA

Os dados coletados que dão sustentação para os trabalhos científicos podem não está disponíveis em um único meio físico. Abbott (2008) salienta que muitas vezes, eles ficam armazenados em PCs e mídias pessoais dos

[15] Dados levantados no repositório institucional ARCA, Fiocruz. 10 out. 2015.

pesquisadores sem que possamos (i) evitar que esse material deixe de ser útil tecnologicamente pela fragilidade das mídias, mas, sobretudo, (ii) por não se pensar na preservação adequada para serem reutilizados em novas pesquisas.

É necessário se pensar que ao longo dos anos se esses dados não estiverem sob cuidados especiais poderão se perder, o que poderá trazer grande prejuízo para a ENSP e para a ciência, uma vez que possam surgir possibilidade de novos investimentos financeiros e de tempo, gerando mais custos nas pesquisas.

Uma das formas para que esses dados sejam armazenados, preservados e acessados hoje e no futuro, é estarem depositados em repositórios de dados e que não haja barreiras de acesso e ainda possam ser compartilhados entre a comunidade científica e a sociedade para uso e reuso.

Atualmente, o grande desafio da ENSP é organizar os dados de projetos de pesquisas concluídos e em andamentos e publicá-los em um repositório de dados científicos. Como projeto piloto, foi escolhida a linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública e Ambiente existente há cerca de 15 anos na escola e pelo apoio da líder de pesquisa e seus pesquisadores em disponibilizar seus dados brutos, objetivando o acesso e o reuso em novas pesquisas científicas.

3 OBJETIVOS

3.1 OBJETIVO GERAL

Organizar as informações dos dados brutos dos trabalhos científicos da linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública e Ambiente produzidos e disponibilizados, nos últimos 15 anos pela ENSP, em um catálogo de dados, de modo a possibilitar o reuso dos mesmos.

3.2 OBJETIVOS ESPECÍFICOS

- Identificar os tipos de dados dos trabalhos científicos da linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública e Ambiente.
- Definir os metadados necessários para representar os conjuntos de dados brutos da pesquisa.
- Organizar os conjuntos de dados brutos em um catálogo utilizando o padrão de metadados definidos no item anterior.

4 DESENVOLVIMENTO

4.1 REVISÃO DA LITERATURA

O crescimento intenso das TIC´S mostra uma variedade de fontes de informação que modificam, ampliam e agilizam a habilidade de comunicação da informação em todo o universo da sociedade. Com isso, caminhos se abrem para favorecer a divulgação de resultados de pesquisas, para o âmbito da ciência, das universidades e instituições correlatas (TOMAEL; SILVA, 2007). A internet surge como uma ponte de circulação de informação, que antes não existia pela dificuldade geográfica, facilitando a divulgação e a recuperação dos trabalhos do campo científico.

Sales e Sayão falam da importância do livre acesso à informação pelos repositórios institucionais, onde afirmam que “os periódicos de acesso livre e os repositórios institucionais vêm se constituindo em alternativas viáveis para que os resultados da pesquisa não pertençam somente ao cientista, e sim à toda humanidade” (SALES; SAYÃO, 2012, p.121).

Da mesma forma, os repositórios institucionais também podem ser vias de divulgação dos dados brutos de pesquisas, que poderão ser reutilizados, acarretando em novas pesquisas e acelerando o processo de geração de novos resultados.

Seguindo esse entendimento, o conceito do acesso livre vai além das publicações tradicionais acadêmicas como os artigos de periódicos e trabalhos acadêmicos de teses e doutorados- um pilar de importância crítica para a prática de uma ciência aberta - não está somente nesses conteúdos, mas além, na disponibilização dos dados de pesquisas (SAYÃO; SALES, 2012).

A ideia de explorar as potencialidades dos dados brutos vem de algumas décadas e mostra que não é uma vontade dos dias atuais. Esta iniciativa não é tão recente e é ressaltada por Sayão e Sales e quando lembram no âmbito da pesquisa sobre primatas[16], o tratamento dos dados que resultou em um catálogo impresso (SAYÃO; SALES, 2012).

[16] Sayão e Sales afirmam que o ”Museu Paraense Emílio Goeldi, em fins da década de 1980 e início de 1990, desenvolveu o PRIMATAM, projeto ligado ao Núcleo de Primatologia, cujo tratamento dos dados de pesquisa resultou em um catálogo impresso.”

O projeto Genoma, famoso por abrir seus dados à comunidade científica e ao público possui um banco de dados chamado Genbank[17] que guarda sequências genéticas de DNA. Este banco está disponível no sítio do *National Center for Biotechnology Information*.

Outras bases de dados surgiram com o movimento de acesso aberto que ampliou as fronteiras dos trabalhos científicos, como a PubMed[18], a PLOS (Public Library of Science)[19], a BioMed Central (BMC)[20], mas com relevância em publicações de artigos, teses e dissertações, sem os dados brutos das pesquisas científicas.

É importante definir o que seriam "dados de pesquisas" ou "dados brutos[21]" ou "dados científico" que são citados em várias bibliografias.

A Organização para a Cooperação Econômica e Desenvolvimento (OCDE) aponta que **dados de pesquisa**:

> [...] dados de pesquisa são definidos como registros factuais (números, registros textuais, imagens e sons) utilizados como fontes primárias para a pesquisa científica, e que são geralmente aceitos na comunidade científica como necessários para validar os resultados da pesquisa (OCDE, 2007, p.13).

No relatório do estado da arte de repositórios de dados científicos de Portugal, Rodrigues *et al.* (2010, p. 48), os definem *dados científicos* como: "dados que são produzidos no contexto de investigação científica ou que de alguma forma são usados em investigação".

Já Abbott (2008) define **dados brutos** como os que servem como fontes de pesquisas.

Junto ao crescimento da ciência devido a investimentos públicos na pesquisa e o avanço das tecnologias da informação e sua capacidade de processamento e de armazenamento de pesquisas científicas, cresce também o volume de dados digitais oriundo dessas pesquisas. Entretanto, dados não estão sendo armazenados e disponibilizados para novas pesquisas. Há uma imensa riqueza de dados de pesquisas que não estão sendo compartilhados. Os dados brutos são muitas vezes esquecidos em mídias pessoais e são perdidos pela

[17] GenBank . Disponível em: <http://www.ncbi.nlm.nih.gov/genbank/> [Consultado em: 24 de agosto de 2015]

[18] http://www.ncbi.nlm.nih.gov/pubmed

[19] https://www.plos.org/

[20] https://www.biomedcentral.com/

[21] **Dados brutos:** são os primeiros dados obtidos.

obsolescência tecnológica e pela fragilidade inerente das mídias digitais (SAYÃO; SALES, 2012).

Na maioria das vezes, estes dados não podem ser coletados novamente e por isso são sensíveis em relação à sua integridade. Caso sejam alterados, destruídos ou perdidos não poderão ser aproveitados futuramente em novas pesquisas. Deve-se considerar que, são muitas as formas e tamanhos dos arquivos de dados nas diversas áreas das pesquisas científicas.

Devemos nos preocupar em mantê-los para assegurar que sejam novamente utilizados em novas pesquisas sem que o formato digital seja obstáculo para uma nova utilização, pois, além de gerar novos dados digitais, os pesquisadores e os acadêmicos, já há algum tempo, começaram creditar toda a confiança nos conteúdos digitais criados por outros cientistas para dar prosseguimento aos seus empreendimentos (SAYÃO; SAL, 2012). Procter, Halfpenny e Voss (2012) alertam que esses dados de pesquisa não terão valor até que sejam gerenciados de forma a assegurar sua recuperação, acessibilidade e reuso.

Para Rodrigues *et al.* (2010), os próprios repositórios institucionais podem ser a solução para que esses dados científicos sejam armazenados e recuperados para uso em outras pesquisas.

Sayão e Sales falam sobre a importância de se estabelecer metodologias e ter comprometimento por um longo prazo, para garantir que a capacidade dos formatos digitais dos dados de pesquisas que estão sendo gerados agora, sejam acessados, interpretados e reutilizados com a tecnologia corrente à época do acesso (SAYÃO; SALES, 2012).

Rodrigues *et al* (2010, p.11) afirmam que para "se constituírem como verdadeiramente úteis, os dados científicos devem possuir estrutura e organização. Os conjuntos de dados (*datasets*) são uma das unidades essenciais". E definem o que seriam conjuntos de dados: "os conjuntos de dados são coleções de informações ou fatos relacionados entre si e registrados num formato comum" (RODRIGUES *et al,* 2010, p.11).

O arquivamento persistente[22], a preservação digital[23], seguido de um modelo de preservação para registros científicos, é a grande questão para a área de pesquisa. Sayão e Sales afirmam que:

> os conhecimentos e as práticas acumulados na última década em preservação digital e acesso resultaram num conjunto de estratégias, abordagens tecnológicas e atividades que agora são coletivamente conhecidas como "curadoria digital (SAYÃO e SALES, 2012, p.184).

À curadoria digital cabe a gestão e a preservação dos recursos digitais para garantir o acesso às gerações atuais e futuras da comunidade científica e sociedade, sem perder a integridade e formatos digitais (SAYÃO e SALES, 2012).

O *Digital Curator Centre* (DCC)[24], é um centro de curadoria digital criado para resolver os desafios da curadoria digital. O DCC (2004) descreve que a curadoria digital "envolve a manutenção, a preservação e a agregação de valor a dados de pesquisa durante o seu ciclo de vida". Possui um modelo para o ciclo de vida dos dados que possui várias ações necessárias para o sucesso do processo de curadoria e de preservação de dados de pesquisa. O Centro propõe uma sequência de ações do modelo de ciclo de vida da curadoria digital: conceituar; criar e receber; avaliar e selecionar; capturar; ação de preservação; armazenar; acessar, usar e reusar; transformar; eliminar; reavaliar; e migrar.

Para a implementação de melhores práticas da curadoria digital, todo processo deve ser de responsabilidades dos envolvidos, isto é, pesquisador, bibliotecário, cientista de dados e gestor de dados. Muitas das decisões tomadas desde o ponto de criação e em várias outras fases do ciclo de vida sofrem impacto nos dados sobre a capacidade de utilização de longa duração de objetos digitais. Portanto, é vital que todos os que lidam com dados da pesquisa compreendam os seus vários papéis e responsabilidades em relação à curadoria digital e preservação de dados. Na figura 1, tem-se a visão geral das funções e responsabilidades dos envolvidos na curadoria digital e preservação.

---

[22] **Arquivo persistente** - Em ciência da computação, **persistência** se refere à característica de um estado que sobrevive ao processo que o criou. Sem essa capacidade, o estado só existiria na memória RAM do computador, e seria perdido quando a RAM parasse (desligando-se o computador por exemplo). Disponível em: https://pt.wikipedia.org/wiki/Persistência_(ciência_da_computação).

[23] **Preservação digital** - é o conjunto de atividades ou processos responsáveis por garantir o acesso contínuo a longo-prazo à informação e a todo patrimônio cultural existente em formatos digitais. Disponível em: https://pt.wikipedia.org/wiki/Preservação_digital

[24] http://www.dcc.ac.uk/

Figura 1 – Habilidades primordiais para gestão de dados

Fonte: Tradução feita pela autora desse trabalho com base em http://www.dcc.ac.uk/

Sayão e Sales (2010, p. 184) reafirmam que "o foco da curadoria digital está na gestão por todo o ciclo de vida do material digital, de forma que ela permaneça continuamente acessível e possa ser recuperado por quem dele precise". Não há como serem recuperados e acessados de forma rápida e fácil se não possuírem os modelos de informações, expressos por metadados, que são ferramentas importantes para que sigam com procedimentos de controle de autenticação (HIGGINS, 2011).

Os usuários esperam que o acesso aos dados científicos seja feito de forma rápida e segura. Barbosa e Sena (2006, p.169) afirmam que, "em geral, a necessidade de localização e acesso rápido a dados específicos, dentro de grandes conjuntos de dados, é comum, tornando relevante a documentação e a organização dos acervos".

Algumas instituições já adotam como solução, o desenvolvimento de catálogos de dados, que presta serviço de localização a análise de dados preliminar de conjuntos de dados (CALLAHAN; JONHSON, 1995). Como definição de catálogo de dados, Barbosa e Sena (2006, p.2) definem que os "catálogos de dados são sistemas de armazenamento que contêm informações descritivas sobre os dados como, por exemplo, seu conteúdo, abrangência temporal e geográfica, e qualidade".

Com o auxílio da internet, viabilizando a disseminação da informação, a tecnologia de banco de dados tem crescido em diversas áreas de aplicações e, com

a utilização de recursos específicos, agrega funcionalidades. Mas, para que dados científicos possam estar em bases de dados com a possibilidade do acesso mundialmente, instituições tem se preocupado com a padronização do conteúdo do que será disponibilizado.

A utilização de um padrão de metadados[25] permite que a definição de uma terminologia para um dado, seja descrita de forma única por diferentes instituições.

Neste sentido, em Londres, no dia 1 de dezembro de 2009, foi fundada a DATACITE[26], uma organização sem fins lucrativos e com o objetivo de estabelecer um acesso mais fácil aos dados de pesquisa na internet, aumentar a aceitação de dados de pesquisa como importante, contribuições citáveis para o registro acadêmico e arquivamento de dados de apoio que irá permitir resultados a serem verificados e propostos para futuros estudos.

A DATACITE oferece vários serviços e ferramentas. Porém, a ideia principal é a citação de dados. Acredita-se que livros e artigos de revistas tem se beneficiado de uma infraestrutura que torna fácil de citar, possuem elementos-chaves no processo de pesquisa, como título e autor. Acreditam que devem citar dados da mesma maneira que citam essas fontes de informação.

A citação de dados pode auxiliar na reutilização e verificação de dados, permitindo maximizar o impacto de dados a serem rastreados e ainda na criação de uma estrutura acadêmica que reconhece e recompensa os produtores de dados.

Além disso, outro serviço importante é a localização de repositórios de dados, através do reg3data.org[27], onde serão depositados os conjuntos de dados. O reg3data.org é um registro global de repositórios de dados de pesquisa que abrange repositórios de dados de pesquisas de diferentes disciplinas acadêmicas. Ele apresenta repositórios para o armazenamento permanente e ao acesso dos conjuntos de dados para pesquisadores, órgãos de financiamento, editores e instituições acadêmicas (REG3DATA.ORG, 2015, tradução da autora). A principal ideia desse site é promover uma cultura de compartilhamento, aumentar o acesso e promover a visibilidade dos dados de pesquisa. Consequentemente, poderá recuperar com facilidade os dados, através dos repositórios de dados que estão basicamente ligados um padrão de metadados.

[25] **Metadados**, ou **Metainformação**, são dados sobre outros dados. Disponível em http://pt.wikipedia.org/wiki/Metadados
[26] https://www.datacite.org/
[27] http://www.re3data.org/

A comunidade científica já aponta a utilização de metadados como solução adequada para garantir os serviços de recuperação mais eficiente e preciso sobre a web (MOURA; CAMPOS, 2002), proporcionando a troca de informações entre integração e fontes digitais heterogêneas. Alguns padrões de metadados foram criados e adaptados para atender as necessidades dos usuários em descrever recursos específicos, são eles: MARC, EAD, TEI, GIRLS, SOIF, Dublin Core (DC), IAFA dentre outros.

Barbosa e Sena (2006, p.4) ressaltam que "o padrão desenvolvido pelo *DublinCore Metadata Initiative* (DCMI) contém um conjunto especializado de expressões para descrição dos recursos eletrônicos a partir da Internet". Entretanto para catalogação de metadados científicos, o DCMI não fornece elementos suficientes, pois os dados podem possuir características particulares que seus elementos não possuem.

Para exemplificar que o padrão de metadados deva ser específico, ressalta-se que o *Government Information Locator Service* (GILS) possui a finalidade de catalogar especificamente informações governamentais. Além desse, há o padrão utilizado para bibliotecas digitais, *Bibliographic*-1 (BIB-1), que serve para cadastrar informações de dados bibliográficos, e o *Geographic Data Commite (*FGDC), específico para dados geo-espaciais, que descreve dados vetoriais e pontuais.

Salienta-se que dada um padrão de metadados específicos, catálogos de dados (CD) é importante como sistemas para informar as descrições dos conjuntos de dados e indicar as suas localizações (BARBOSA; SENA, 2006), tem como fator chave a análise dos dados e a possibilidade dos usuários determinarem, se querem adquiri-los para uso de novas pesquisas científicas.

A fim de que os catálogos de dados sejam utilizados fortemente é preciso que as instituições documentem todos os conjuntos de dados de seus acervos. De acordo com Callahan e Johnson (1995), seis fatores chaves devem ser considerados durante o processo de desenvolvimento desses sistemas (CD): a *Completude* para que as instituições façam a documentação de todos os conjuntos de dados de seus acervos; a *Facilidade* de utilização a fim de se promover treinamentos extensivos para a utilização de um sistema; a *Coerência* das informações para que o conteúdo dos CD deve ser determinado criteriosamente por quem classifica os dados, uma vez que sua utilidade depende da relevância das informações que são retornadas pelas consultas; a *Precisão* está relacionada ao fato de que o CD deve ser preciso e evitar

descrições incompletas, que podem acarretar baixa credibilidade; pois muitos usuários acessam um CD para teste de veracidade das informações armazenadas; a *Disponibilidade* é a facilidade de acesso a partir das redes de computadores que deve permitir que qualquer usuário dentro das instituições possa ter acesso às informações dos CD; e por fim devem *Serem Públicos*, pois as pessoas devem saber que estes sistemas existem, entender que devem ser utilizados e aplicá-los em beneficio do desenvolvimento de seus trabalhos.

Barbosa e Sena (2006, p.173) explicam que "um dos objetivos destes sistemas é viabilizar o acesso e a localização dos conjuntos de dados de maneira rápida e fácil". Os conjuntos de dados possuem diferentes tipos de informação e, com isso, os CD devem ser flexíveis para suportar as variações.

Uma mudança de conceitos, reestruturação organizacional, aprendizagem e planejamento nas instituições, são itens que incorporam no processo de descrição ou catalogação de dados, não é apenas uma questão tecnológica. Esse processo, no que diz respeito a dados científicos, é a importância dada na utilização de metadados para documentar. A partir dessas questões, possibilitará disponibilizar e ampliar novos estudos científicos com compartilhamento e acesso livre a toda comunidade científica.

5 METODOLOGIA

Segundo a classificação proposta por Tobar e Yalour (2004) a metodologia proposta é de cunho qualitativo e consistirá na realização de uma pesquisa de natureza exploratória. A escolha por essa metodologia se deve ao fato da inexistência de estudos sobre este tema abordando a instituição escolhida.

Sendo assim, de sorte a atender ao objetivo de **identificar os tipos de dados dos trabalhos científicos da linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública e Ambiente** será necessário analisar o conteúdo dos trabalhos científicos, optando-se pela análise categorial ou temática, na tentativa de identificar e classificar sobre a relação semântica dos dados coletados.

A análise categorial ou temática é descrita pela autora Oliveira (2008) como sendo o que permite a exploração do material analisado a partir da observação de diferentes elementos presentes no texto, bem como conduzem a resultados distintos em termos de compreensão da mensagem.

Inicialmente será feito o levantamento dos trabalhos científicos referentes aos pesquisadores da linha de pesquisa Gestão em Saúde e Ambiente da Escola de Saúde Pública Sérgio Arouca. A relação dos pesquisadores dessa linha de pesquisa será capturada através do acesso a plataforma Lattes cujo a consulta a base será pela parametrização do nome da linha de pesquisa e pesquisadores.

De posse da lista com os nomes dos pesquisadores vinculados à referida linha de pesquisa, serão identificados, inicialmente, os trabalhos científicos depositados no repositório institucional (RI) da Fiocruz (ARCA).

A escolha do RI ARCA se dá ao fato da instituição possuir uma política mandatória de depósito em bases de acesso aberto das pesquisas de pesquisadores da Fiocruz cujo o investimento recebido para o desenvolvimento das mesmas é originado de fontes públicas de financiamento.

Outra base utilizada será o repositório temático ENSP, pois esta Unidade possui os artigos científicos publicados por pesquisadores "da casa" utilizando o procedimento de auto arquivamento.

Embora PUBMED[28], Scielo[29] e BVS[30] abriguem uma série de periódicos de acesso aberto e de grande importância, por este trabalho se tratar de um projeto piloto, estes não serão consultados.

De posse deste levantamento, serão armazenados os trabalhos científicos em uma planilha Excel, sem duplicidades.

Em seguida, na lista será analisada o contexto da pesquisa, de acordo com a semântica, e classificados em grupos de trabalhos científicos a partir da leitura dos resumos e palavras chaves.

Após a classificação dentro dos grupos listados acima, os artigos serão lidos integralmente com o objetivo de identificar quais os tipos de dados que serviram como instrumentos científicos, a partir da análise do conteúdo dos documentos disponibilizados.

Para atingir o segundo objetivo específico de **definir os metadados necessários para representar os dados brutos da pesquisa** será realizado, inicialmente, a identificação dos metadados utilizados pelo registro de repositórios de dados de pesquisa – reg3data.org[31], em três repositórios de dados brutos: Harvard

[28] Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>
[29] Disponível em: <http://www.scielo.org/php/index.php>
[30] http://bvsalud.org/
[31] http://www.re3data.org/

Dataverse, Washington State University Data Center Dataverse e Russia Longitudinal Monitoring Survey Dataverse, verificando similaridades semântica destes metadados, ou seja, se o significado do metadado em um repositório possui o mesmo significado em outro repositório. Será criada uma tabela contendo o nome do metadado e seu significado.

Entretanto, para verificar quais destes metadados são pertinentes para esta linha de pesquisa, buscar-se-á agrupar os tipos de dados encontrados e relacioná-los com os metadados descritos na tabela.

Por fim, **organizar os conjuntos de dados no catálogo de dados utilizando o padrão de metadados definidos no item anterior,** que será construído na ferramenta open source WordPess ou Joomla - Gerenciador de conteúdo, por ser software livre e atende requisitos de arquitetura de informação organizada por categorias, que facilitará buscas parametrizadas. Os conjuntos de dados disponibilizados estarão nos formatos de arquivos: CSV, HTML, TXT,PDF, XLS, entre outros. E ainda, acompanhados com metadados associados e acesso de links de URL, que estarão disponibilizados em um repositório de dados científico oferecido pelo reg3data.org.

Para recuperar os conjuntos de dados, o usuário poderá optar por filtros de pesquisas nos vários tipos disponíveis para buscas, tais como: tópicos, categorias, tipo de conjunto de dados, tags, formatos, tipo de organização e autor. Dessa forma, poderão consultar os conjuntos de dados com várias informações agregadas acompanhados com os metadados definidos.

Ao acessar, o conjunto de dados escolhido, será direcionado ao repositório de dados científico de acesso livre e poderá obter informações da descrição do conjunto de dados, acesso e uso da informação, recursos (dados) e download e informações de metadados.

O sistema de Catálogo de dados estará disponível no Portal da ENSP a fim de divulgar os dados brutos, a localização para o acesso, uso e reuso e servirá ainda, de estímulo a promoção científica dos dados brutos de pesquisas para as outras linhas de pesquisas da própria ENSP, e à outras unidades da Fiocruz.

Essa proposta piloto de catálogo de dados foi inspirado no site A casa de dados aberto do Governo EUA[32].

[32] http://www.data.gov/

6 CONSIDERAÇÕES PARCIAIS/FINAIS

Considerando a ENSP uma unidade da Fiocruz que atua em pesquisa científica na área de saúde pública e que busca ampliar o acesso pleno de seu conhecimento a toda sociedade, espera-se que este projeto (i) amplie as possibilidades de acesso aos dados já coletados, (ii) diminua o gasto com coletas de dados que já tenham sido coletados, e (iii) diminua o tempo nas coletas de dados para pesquisas complementares às já realizadas. Para isso, o catálogo proporcionará a ENSP oferecer à comunidade científica interna e outras instituições, quais grupos de dados brutos identificados e classificados levantados nas pesquisas **dos trabalhos científicos da linha de pesquisa Gestão Ambiental e Saúde do programa Saúde Pública e Ambiente** possam contribuir com o desenvolvimento de novas pesquisas que tenham interesse nessa temática.

REFERÊNCIAS

ABBOTT, D. "**What is digital curation?**". DCC Briefing Papers: Introduction to Curation. Edinburgh,2008: Digital Curation Centre. Handle: 1842/3362. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation>. Acesso em: 31 ago. 2015.

BAPTISTA, Ana Alice; COSTA, Sely M. S.; KURAMOTO, Hélio; RODRIGUES, Eloy. Comunicação científica: o papel da Open Archives Initiative no contexto do acesso livre. **Encontros Bibli**: Revista Eletrônica Biblioteconomia Ciência da Informação, Florianópolis, n. esp., 1. sem. 2007. Acesso em: 7 out. 2015.

BARBOSA, E.B.M.; SENA, G. J. de. Um banco de metadados para auxiliar a disseminação de dados científicos em instituições de pesquisas. In: CONGRESSO INTERNACIONAL DE GESTÃO DA TECNOLOGIA E SISTEMAS DE INFORMAÇÃO,4., 2006, São Paulo. **Anais...** São Paulo: CONTEGSI,3. Disponível em: http://www.contecsi.fea.usp.br/envio/index.php/contecsi/3contecsi/paper/download/2073/1177. Acesso em: 17 set. 2015.

BERLIN Declaration on Open Access to Knowledge in the Sciences and Humanities. Berlin, 2003. Disponível em:< http://openaccess.mpg.de/Berlin-Declaration>. Acesso em: 18 ago. 2015.

CALLAHAN, S.D.; JONHSON, B.D. Scientific data set catalogues. In: AGSO FORUM ON GIS IN THE GEOSCIENCES, 2nd., 1995, Canberra. **Proceedings...** Canberra: ACT, 1995. p. 29-31. Acesso em: 17 set. 2015.

CHALHUB, Tania; BENCHIMOL, Alegria; GUERRA, Claudia. Acesso livre via repositórios: políticas de instituições brasileiras. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, p. 159-173, dez. 2012. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/15182924.2012v17nesp2p159>. Acesso em: 21 set. 2015.

HIGGINS, S. Digital Curation: the emergence of a new discipline. **The International Journal of Digital Curation**, v.6, n. 2, 2011. Disponível em: <http://www.ijdc.net/index.php/ijdc/article/view/184> Acesso em: 17 set.2015.

KURAMOTO, Hélio. Acesso livre à informação científica: novos desafios. **Liinc em revista**, v. 4, n. 2, p. 155-158, 2008.
_____. Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica. 2008. Disponível em: http://kuramoto.files.wordpress.com/2008/09/ manifesto-sobre-o-acesso-livre-a-informacao-cientifica.pdf Acesso em: 24 ago. 2015.

LEITE, F. C. L. et al. **Como gerenciar e ampliar a visibilidade da informação científica brasileira**: repositórios institucionais de acesso aberto. Brasília: IBICT, 2009. Disponível em: < http://livroaberto.ibict.br/bitstream/1/775/4/Como%20gerenciar%20e%20ampliar%20a%20visibilidade%20da%20informa%C3%A7%C3%A3o%20cient%C3%ADfica%20brasileira.pdf>. Acesso em: 24 ago. 2016.

MOURA, A. M. C.; CAMPOS, M. L. M. A Metadata approach to manage and organize electronic documents and collections on the web. **Journal of the Brazilian Computer Society**, v. 1, n. 8, p. 16, 2002.

MEDRI, W. **Análise exploratória de dados**. 2011. p.13. Disponível em: <http://www.uel.br/pos/estatisticaquantitativa/textos_didaticos/especializacao_estatistica.pdf>. Acesso em: 20 ago. 2015.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Principles and guidelines for access to research data from public funding.** Paris: OCDE, 2007. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdf>. Acesso em: 15 set. 2015.

OLIVEIRA, D. C. **Análise de conteúdo temática**: uma proposta de operacionalização: texto didático e instrumentos. Rio de Janeiro: Universidade do Estado do Rio de Janeiro, 2004.

POLÍTICA e repositório da Ensp obtêm registro internacional. Disponível em: <http://www.ensp.fiocruz.br/portal-ensp/informe/site/materia/detalhe/32249>. Acesso em: 18 nov. 2015.

PROCTER, R.; HALFPENNY, P.; VOSS, A. Research data management: opportunities and challenges for HEIs. In: PRYOR, Graham (Org.). **Managing research data**. Londres: Facet Publishing, 2012. Chapter 7, p. 135-150.

SAYAO, L. F.; SALES, L. F. **Curadoria digital**: um novo patamar para preservação de dados digitais de pesquisa, 2012. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/12224/8586>. Acesso em: 17 set. 2015.

SALES, Luana Farias; SAYÃO, Luís Fernando. O impacto da curadoria digital dos dados de pesquisa na Comunicação Científica. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, p. 118-135, dez. 2012. ISSN 1518-2924. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/15182924.2012v17nesp2p118>. Acesso em: 24 ago. 2015.

ENCONTRO Nacional de Pesquisa em Ciência da Informação, 14.,2013, Florianópolis. GT 7: produção e comunicação da informação em ct&i. Disponível em <http://enancib.ibict.br/index.php/enancib/xivenancib/paper/ viewFile/4347/3470>. Acesso em: 14 set. 2015.

YAMAOKA, E. J. Ontologia para mapeamento da dependência tecnológica de objetos digitais no contexto da curadoria e preservação digital. **AtoZ**, Curitiba, v. 1, n. 2, p. 65-78, jan./dez. 2012.

SEMINÁRIO Internacional de Acesso Livre ao Conhecimento: impactos na produção acadêmica, divulgação científica e inovação no ensino, 1, 2011. Disponível em: <http://www.ensp.fiocruz.br/portal-ensp/informe/site/evento/detalhe/14982>. Acesso em: 18 nov. 2015.

TOBAR, F.; YALOUR, M. **Como fazer teses em saúde pública**. Rio de Janeiro: Fiocruz, 2004.

TOMAÉ, M. I.; SILVA, T. E. **VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação**: 28 a 31 de outubro de 2007 • Salvador • Bahia • Brasil. Disponível em <http://www.enancib.ppgci.ufba.br/artigos/GT5--142.pdf>. Acesso em: 15 set. 2015.