

Processo de indexação para repositórios institucionais: Introdução ao contexto

Mariângela Spotti Lopes Fujita

Pesquisadora Produtividade em Pesquisa
do CNPq - Nível 1-B

Docente Permanente do PPGCI/UNESP -
Campus de Marília



Repositório Institucional

O repositório institucional é definido como “[...] sistemas de informação que armazenam, preservam, divulgam e dão acesso à população intelectual de comunidades científicas [...]” (IBICT, 2015).



Os repositórios digitais em universidades são, atualmente, recursos de grande relevância na organização e gestão do conhecimento gerado pela produção científica, tecnológica, artística e administrativa (Fujita, 2018).

REPOSITÓRIOS INSTITUCIONAIS

Sistemas de recuperação da informação pensados para atender aos modos contemporâneos de criar conhecimento e gerir informações de acesso aberto.

Exigem formas inovadoras de se tratar, recuperar e acessar dados e informações oriundos de uma crescente produção tecnocientífica, frente às exigências e especificidades do acesso aberto.

Carecem de políticas que atendam à complexidade de um sistema global e que considere todas as etapas do processo informacional em graus igualmente expressivos.

REPOSITÓRIOS INSTITUCIONAIS: características

O contexto de repositórios, tanto de gestão quanto de organização e representação da informação, tem características diferentes das de outros sistemas de armazenamento e recuperação da informação, cujas decisões políticas não se aplicam.

Exemplos dessa diferença são:

- uso de linguagem natural por palavras-chave,
- a migração de grandes quantidades de dados, arquivos e registros bibliográficos de outros sistemas,
- Diferentes tipos de armazenamento.

Tipos de armazenagem no Repositório

- ❖ Autoarquivamento: Consiste no processo em que o próprio autor torna-se responsável pelo depósito de seu trabalho no Repositório. O processo possibilita que o depositante cadastre identificadores e descritores do material submetido.
- ❖ Depósitos Mediados: São submissões de documentos realizadas no repositório através de terceiros, normalmente bibliotecários ou parceiros autorizados pelo autor do material)
- ❖ Marcação (Tagging): Tagging é a atribuição livre e pessoal de palavras para descrever ou apontar o conteúdo que está sendo inserido durante o autoarquivamento.
- ❖ Povoamento: Coleta, captura ou colheita de registros em bases de dados feita de forma automática ou semiautomática a fim de trazer os documentos produzidos por pesquisadores para o repositório da sua instituição.



Funcionamento do repositório

Depende de políticas internas da instituição orientadas por diretrizes, padrões e normativas pré-existentes que visam a organização, representação e recuperação.

A equipe de profissionais executará atividades de gestão e de organização e representação da informação específicas do contexto e da cultura informacional do repositório.

A política, portanto, é formulada pela equipe de profissionais considerando o contexto institucional dotado de requisitos de infraestrutura, comunidade de usuários e, sobretudo objetivos e interesses institucionais.

Políticas dependem do estudo das atividades organizacional, de gestão e da organização e representação da informação, porque são característicos de cada contexto em que se inserem, e nenhum contexto será igual ao outro.

Formulação da política de indexação em repositórios



Conhecer os envolvidos na formulação da política de indexação:



Quem são os atores que realizam a gestão do repositório institucional? (**Gestores**: bibliotecários e outros profissionais);



Quem são os atores que executam as atividades de organização e representação? (**Executores**: catalogadores e criadores de registros bibliográficos);



Quem são os atores que consultam os registros bibliográficos do repositório, em suas buscas? (**Usuários**: pesquisadores e administradores institucionais).

Formulação da política de indexação em repositórios

Análise do funcionamento e avaliação da recuperação da informação: a visão abrangente: coleta de dados por observação, conhecimento da documentação, aplicação de questionário com entrevistas, avaliação do sistema de recuperação da informação;

Definição dos requisitos, elementos e variáveis da política de organização e representação da informação para repositórios

ATIVIDADES DE ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO: **ELEMENTOS**

1. Cobertura de assuntos
2. Seleção e coleta por povoamento, migração ou autoarquivamento de documentos-fonte
3. Processo de descrição física
4. Processo de indexação
5. Estratégia de busca
6. Forma de saída
7. Avaliação do sistema

Processo de Indexação

É um processo de análise de assunto para representação do conteúdo significativo de documentos e recursos informacionais digitais ou impressos;

O objetivo final da indexação é o acesso e recuperação por assuntos de forma específica e/ou exhaustiva com e sem controle de vocabulário em sistemas de informação.

A determinação de assuntos, seja por palavra-chave ou descritor, seja por humanos ou por máquina, é indexação.

Variáveis do processo de indexação

Nível de exaustividade

Nível de especificidade

Linguagem de indexação versus linguagem livre

Recuperação da informação

- ❖ **Especificidade:** é um princípio de qualidade do processo de indexação com o objetivo de garantir que o termo de indexação a ser extraído do documento seja o de nível mais específico em uma hierarquia de termos gerais. Lancaster (2004, p.34), explica o princípio de especificidade, com o seguinte exemplo: “[...] um artigo que trate do cultivo de laranjas será indexado sob LARANJAS e não sob FRUTAS CÍTRICAS ou FRUTAS.” Neste exemplo o termo LARANJAS é o termo de nível mais específico da hierarquia, diretamente subordinado ao termo FRUTAS CÍTRICAS; e o termo mais geral é FRUTAS.
- ❖ **Exaustividade:** A exaustividade, entretanto, é o princípio que garante a existência de especificidade porque permite o “emprego de termos em número suficiente para abranger o conteúdo temático principal do documento. Quanto mais termos forem utilizados para indexar um documento, mais acessível ele se tornará e, provavelmente, mais vezes será recuperado” (LANCASTER, 2004, p. 27).
- ❖ **Correção:** ausência de erros de inclusão e de omissão
- ❖ **Consistência:** grau de coincidência entre duas ou mais indexações

O processo de indexação

“Princípios de indexação”
e a Norma ABNT 12.676

ANALÍTICO (análise de assunto):

- Compreensão do texto como um todo,
- Identificação de conceitos,
- Seleção de conceitos

TRADUÇÃO:

Representação de conceitos por termos de um vocabulário controlado → controle de vocabulário

Controle de vocabulário e Vocabulário controlado

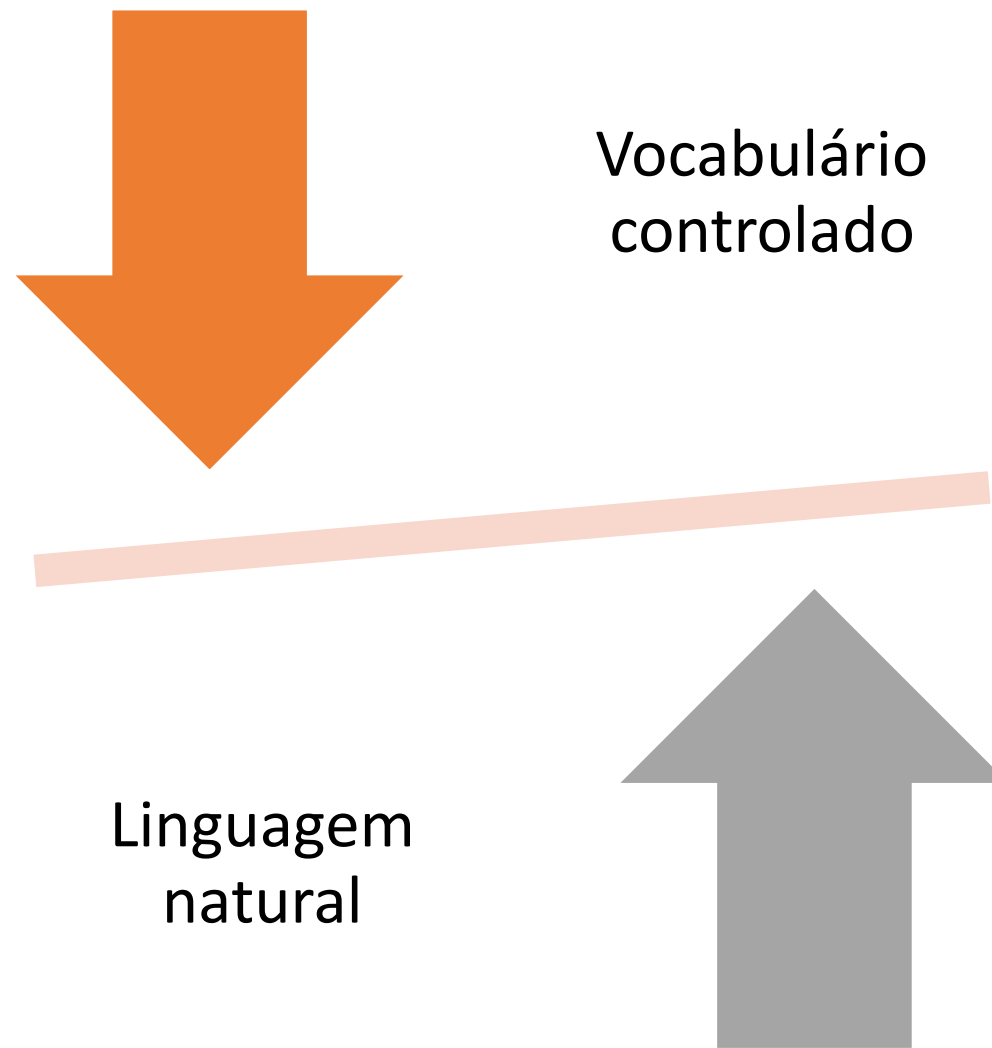
Controle de vocabulário

“O principal objetivo do controle de vocabulário é obter consistência na descrição de objetos de conteúdo e facilitar a recuperação.” (ZENG, 2005, p. 1).

Vocabulário controlado

Lista enumerada de termos controlados que busca eliminar a ambiguidade, redundância e promover o controle dos sinônimos. Possui objetivo de promover a organização da informação, na qual a atribuição de termos retirados de algum vocabulário controlado é utilizada para descrever o conteúdo de documentos ou outros objetos. Podem estar estruturados em: listas, anéis de sinônimos, taxonomias e tesauros (ANSI/NISO Z39.19:2005).

O processo de indexação e o controle de vocabulário em repositórios → recuperação da informação



PROBLEMAS

A falta de controle de vocabulário pode gerar vários tipos de inconsistências determinantes para a imprecisão e irrelevância na recuperação de informações em qualquer sistema de busca (Fujita; Tolare, 2019).

- Polissemia.
- Falta de visibilidade das publicações;
- Grafias diferentes de um mesmo conceito;
- Singular e plural

O CONTROLE DE VOCABULÁRIO NA INDEXAÇÃO E NA RECUPERAÇÃO DA INFORMAÇÃO

VANTAGENS

- Precisão terminológica do domínio especializado;
- Controle da sinonímia, homonímia e polissemia;
- Correção da escrita de termos;
- Controle de singular e plural;
- Recuperação com precisão e revocação
- Adição de itens bibliográficos semelhantes sem dispersão

DESVANTAGENS

- Diminuição da especificidade em favor da exaustividade;
- Tendência de uso de termos mais gerais;
- Atualização lenta de vocabulário;
- Falta de exatidão no uso de termos sinônimos;
- Incapacidade de cobertura do domínio especializado

Como os repositórios institucionais podem melhorar a recuperação de informações pelos usuários, de modo a aumentar a compatibilização entre a forma de indexar e a forma de buscar (recuperar) informações?



O CONTROLE DE VOCABULÁRIO NA INDEXAÇÃO E NA RECUPERAÇÃO DA INFORMAÇÃO: recomendações

Desejável a integração das modalidades de vocabulário controlado e linguagem natural nos campos de assunto dos metadados para o aumento de possibilidades de acesso.

Usuários que realizam autoarquivamento poderiam ter acesso ao vocabulário controlado como mais uma opção de representação dos assuntos dos documentos e ampliação da garantia de recuperação.

Aprimoramento e aumento das palavras-chaves a partir de uma fonte terminológica positiva e confiável da sua área de especialidade.



Ciência aberta e repositórios

Representação da Informação

Abordagens da representação da informação (CHU, 2010):

Indexação com diversos atores e máquinas: livre ou controlada; manual ou automática; ambientes de hiperestrutura na web; marcação social

Categorização: representações hierárquicas de informação por categorias; classificações hierárquicas, taxonomias, folksonomias e ontologias

Sumarização: resumos, sumários, extratos

Outros métodos: citações (meios automáticos) e indexação em cadeias de termos para criação de resumos indicativos

Representação da Informação

Tópicos relacionados (CHU, 2010)

1 Metadados:

- Padrões de metadados: AACR, MARC (criados antes da era digital) Dublin Core, Resource Description Framework (RDF) para a web; Sistema DOI (Digital Object Identifier) localizador identificador para qualquer objeto digital na web;

2 Texto completo: dificuldades na representação

3 Informação multimídia: som, texto derivado de som e imagens e imagens

Palavras-chave na representação e recuperação de informações em acesso aberto

- O emprego da palavra-chave vai além da busca e passa a ter aplicação na indexação, recuperação de informações, marcação social, extração de palavras-chave, bibliometria e desenvolvimento de tesouros e outros sistemas de organização do conhecimento (LU; LI; ZHIFENG; CHENG, 2019, p. 415).
- A relevância de palavras-chave para tais aplicações reside no fato de que é o próprio autor quem garante a representatividade “chave” de seus textos.
- O autor, quando atribui palavras-chaves, preenche um metadado, sem orientação ou auxílio profissional, que será preservado quanto à padronizações ou qualquer proposta de controle de vocabulário, após preenchimento.
- o autor, ao atribuir palavras-chave torna-se um indexador e precisa pensar em outras funções além da representação do conteúdo do texto e adotar padrões que sejam compatíveis com diferentes empregos futuros.
- as maiores inconsistências estão ligadas à escolha em português e/ou inglês, à adoção de singular ou plural, erros de grafia, palavras-chave em caixa alta ou em caixa baixa, em sua totalidade etc.

Principais desafios e recomendações para a representação e recuperação da informação

Sobre a falta de controle de vocabulário em acesso aberto:

A falta de controle de vocabulário gera inconsistências determinantes para a imprecisão e irrelevância em qualquer sistema de busca na web.

É preciso disponibilizar o vocabulário controlado para consulta durante buscas e tutoriais de como se realiza o controle de vocabulário.

Para o controle de vocabulário é necessária a construção e contínua atualização de vocabulários controlados utilizando a linguagem natural das palavras-chave e fonte terminológica positiva e confiável da área de especialidade.

Recomendação do estudo mais detalhado do controle de vocabulário em listas de palavras-chave em ordem alfabética.

Principais desafios e recomendações para a representação e recuperação da informação

Sobre o autoarquivamento em acesso aberto

Usuários que realizam autoarquivamento poderiam ter acesso ao vocabulário controlado como mais uma opção de representação dos assuntos dos documentos e ampliação da garantia de recuperação.

É necessário avançar na concepção de indexadores não proficientes e elaborar proposta de política de indexação para padronização de palavras-chave atribuídas por autores e pesquisadores, na submissão da produção científica em diferentes sistemas de informação que realizam a gestão e divulgação científica.

Principais desafios e recomendações para a representação e recuperação da informação

Sobre os metadados em acesso aberto

Para facilitar a recuperação e interoperabilidade, é necessário pensar na representação dos documentos por metadados conforme padrões.

Desejável a integração das modalidades de vocabulário controlado e linguagem natural nos campos de assunto dos metadados para o aumento de possibilidades de acesso.

Principais desafios da representação e recuperação da informação

- A solução de problemas de representação e recuperação deve ser estudada, após avaliações de diferentes perspectivas, conforme características de cada sistema de armazenagem e recuperação da informação
- Diferença entre linguagem natural e controle de vocabulário: principal debate da área de Organização e Representação do conhecimento.
- Repositórios institucionais de acesso aberto são utilizados muito mais para fins de preservação digital dos recursos informacionais provenientes da instituição do que como ferramenta de recuperação da informação.
- Sistemas de acesso aberto utilizam softwares que não incluem as padronizações desenhadas para a representação da informação (processos e sistemas de organização do conhecimento) cujos avanços otimizam a recuperação da informação – são diferentes, porém são as bibliotecas e os bibliotecários quem os administra junto com os demais sistemas tradicionais

Referências

AMERICAN NATIONAL STANDARDS INSTITUTE/ NATIONAL INFORMATION STANDARDS ORGANIZATION. Z39.19-2005. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Bethesda, Maryland: NISO Press, 2005. Disponível em: <http://www.niso.org/standards/resources/Z39-19-2005.pdf> . Acesso em: 17 set. 2024

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 12676: Métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação*. Rio de Janeiro, 1992. 4 p.

FUJITA, M. S. L. A política de indexação para representação e recuperação da informação In: FUJITA, M. S. L; GIL LEIVA, I. *Política de indexação*. 1 ed. São Paulo; Marília: Cultura Acadêmica; Oficina Universitária, 2012, v.1, p. 17-30.

FUJITA, M. S. L. Controle de vocabulário na representação e recuperação da informação em repositórios universitários. 2018. *Projeto de Pesquisa* (Bolsa de Produtividade em Pesquisa) - CNPq, Programa de Pós-Graduação em Ciência da Informação, Unesp, Marília, 2018.

FUJITA, M. S. L.; TOLARE, J. . Análise do controle de vocabulário em repositórios institucionais brasileiros. In: Thiago Henrique Bragato Barros; Natalia Bolfarini Tognoli. (Org.). *Organização do conhecimento responsável: promovendo sociedades democráticas e inclusivas*. Belém: Editora da UFPA, 2019, v. 5, p. 432-444.

GIL LEIVA, I. *Manual de indización: teoría e práctica*. Gijón: Trea, 2008.

INSTITUTO BRASILEIRO DE CIÊNCIA E TECNOLOGIA - IBICT, 2015. Disponível em: <http://www.ibict.br/pesquisa-desenvolvimento-tecnologico-e-inovacao/Sistema-para-Construcao-de-Repositorios-Institucionais-Digitais>. Acesso em 17 set. 2024.

LANCASTER, F. W. El control del vocabulario en la recuperación de información. València: Universitat de València, 2004. p.159. (Educació. Materials, 12).

WORLD INFORMATION SYSTEM FOR SCIENCE AND TECHNOLOGY. Princípios de indexação. *R. Esc. Bibliotecon. UFMG*, v.10, n.1, p.83-94, 1981

ZENG, M. L. Construction of controlled vocabularies: a primer (based on Z39.19). 2005.

Livros para uso e leitura

- [FUJITA, M. S. L.](#). A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias. 1. ed. São Paulo: Editora UNESP, 2009. v. 1. 104p.
- [GIL LEIVA, I.](#) (Org.) ; [FUJITA, M. S. L.](#) (Org.) . Política de indexação. 1. ed. São Paulo: Cultura Acadêmica; Oficina Universitária, 2012. v. 1. 260p.
- [FUJITA, MARIÂNGELA SPOTTI LOPES](#); Moreira, Walter (Org.) . Manual do planejamento, construção e manutenção do tesauro Unesp para bibliotecas: do conceitual a prática. 1. ed. Marília; São Paulo: Faculdade de Filosofia e Ciências, 2021. v. 1. 222p.
- [FUJITA, M. S. L.](#); ALVES, R. V. (Org.) ; ALMEIDA, C. C. (Org.) . Modelos de leitura documentária para indexação : abordagens teóricas interdisciplinares e aplicações em diferentes tipos de documentos. 1. ed. Marília; São Paulo: Oficina Universitária; Cultura Acadêmica, 2020. v. 1. 460p.
- [FUJITA, M. S. L.](#); [NEVES, D. A. B.](#) (Org.) ; DAL' EVEDOVE, P. R. (Org.) . Leitura documentária: estudos avançados para a indexação. 1. ed. Marília; São Paulo: Oficina Universitária; Cultura Acadêmica, 2017. v. 1. 318p.
- [FUJITA, M. S. L.](#). Política de indexação para bibliotecas: elaboração, avaliação e implantação. 1. ed. Marília; São Paulo: Oficina Universitária; Cultura Acadêmica, 2016. v. 1. 142p.