



# Metadados descritivos em acervos digitais: um estudo de caso no Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi

**Stela Andrade Vasconcelos<sup>I</sup>**

<sup>I</sup> Universidade Federal do Pará, Belém, PA, Brasil;  
stelavasncelos@gmail.com; <https://orcid.org/0000-0003-1154-3089>

**Ana Vilacy Galucio<sup>II</sup>**

<sup>II</sup> Museu Paraense Emílio Goeldi, Belém, PA, Brasil;  
avilacy@museu-goeldi.br; <https://orcid.org/0000-0003-0168-1904>

**Resumo:** Com os avanços técnicos-tecnológicos recentes na área de informação, há uma proliferação de arquivos digitais, de modo que é importante haver um controle de informações em bases de dados digitais. A utilização de um sistema padronizado de metadados pode ser uma solução para a problemática da dispersão de informações, uma vez que o padrão de metadados gerencia e organiza dados segundo a sua funcionalidade. Este artigo tem como objetivo analisar a criação e a padronização de metadados do Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi. Neste estudo, busca-se aplicar os princípios da Teoria Geral da Terminologia, Teoria Comunicacional da Terminologia e da Teoria da Indexação interligando-os às áreas da Linguística e da Ciência da Informação. A metodologia consistiu na análise dos tipos de recursos informacionais presentes no acervo de línguas indígenas do Museu Paraense Emílio Goeldi e no desenvolvimento e aplicação de uma política de catalogação/indexação na criação dos metadados e padronização dos metadados já existentes e dos novos materiais a serem incluídos nesse acervo. Os resultados iniciais demonstraram que a criação dos metadados condizentes com o conteúdo dos arquivos de mídia possibilita a identificação prévia do recurso informacional. O processo de criação de metadados descritivos pode ser padronizado por meio de uma política de indexação ou manual sobre a criação de metadados, de forma a fornecer bases satisfatórias para as necessidades informacionais dos usuários.

**Palavras-chave:** metadados; terminologia; indexação; recuperação da informação

## 1 Introdução

A terminologia enquanto disciplina estuda a palavra em um contexto formal, em áreas especializadas. Na área dos estudos terminológicos, os termos são usados

para designar entidades do conhecimento produzido em um determinado domínio. Com os avanços técnicos-tecnológicos recentes, há uma proliferação de arquivos digitais, de modo que é importante haver um controle de informações em bases de dados digitais.

A utilização de um sistema padronizado de metadados pode ser uma solução para a problemática da dispersão de informações, uma vez que o padrão de metadados gerencia e organiza dados segundo a sua funcionalidade. Riley (2017) aponta para quatro tipos de metadados: descritivo, administrativo, estrutural e codificador (linguagem de marcação)<sup>1</sup>. Os metadados descritivos, por exemplo, são usados como identificador de um determinado recurso informacional e permitem que o mesmo seja recuperado pelo indivíduo a partir do dado que o descreve.

Em um sistema de recuperação da informação, a interação entre usuário e sistema é fundamental, pois o usuário terá que utilizar uma estratégia de busca eficiente para que o sistema recupere a informação desejada. Por isso, os metadados precisam estar padronizados e coerentes com o recurso informacional indexado, a fim de que os instrumentos de padronização do vocabulário possam auxiliar no controle de uso dos termos.

Como um estudo de um caso específico, este artigo tem como objetivo analisar a criação e a padronização de metadados do Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (ALIM/MPEG) (MPEG, 2023). Como problema de pesquisa, indaga-se como otimizar a criação de metadados utilizando as teorias da terminologia e indexação? Além disso, para avaliar a eficiência dos metadados criados e inseridos no acervo, foi feita uma comparação entre os metadados da língua Aikanã, Makurap e Wayoro, e num segundo momento, a análise da recuperação da informação no acervo digital, tendo duas coleções de línguas indígenas como exemplo a Makurap e a Wayoro.

Neste estudo, busca-se aplicar os princípios da Teoria Geral da Terminologia (Wüster, 2010), Teoria Comunicacional da Terminologia (Cabré, 1993) e da Teoria da Indexação (Lancaster, 2004), interligando-os às áreas da Linguística e da Ciência da Informação, para a criação de um padrão de metadados do Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi

(MPEG, 2023). O restante deste artigo está organizado da seguinte forma. A seção dois apresenta as teorias da Terminologia e da Indexação, com foco específico sobre a Teoria Geral da Terminologia de Eugen Wüster (Wüster, 2010), e a Teoria Comunicativa da Terminologia de Maria Teresa Cabré (Cabré, 1993).

Adiante, na seção três, discutem-se os conceitos de metadados segundo a literatura de Ciência da Informação, e os tipos de metadados listados por Felipe Arakaki (2019). Na seção quatro, a metodologia é apresentada e detalhada por meio de sete etapas. Os resultados da comparação de metadados e da recuperação da informação na base de dados são apresentados na seção cinco. Por fim, as considerações finais retomam o uso das teorias para a criação de metadados e a correlação entre as áreas da Linguística e Ciência da Informação, bem como a contribuição dos metadados para a preservação de recursos informacionais de um acervo digital.

## **2 Teorias da terminologia e da indexação**

A ciência da terminologia surge em torno de 1930, como uma área voltada para a criação de sistemas de termos (Santamaría-Pérez, 2009). Essa área de estudos foi se desenvolvendo ao longo dos anos e passou por vários períodos e fases no seu desenvolvimento histórico, indo desde um período de revisão sobre o papel da terminologia até o uso da informática para facilitar o trabalho dessa área (Santamaría-Pérez, 2009).

Na literatura, considera-se o austríaco Eugen Wüster como o pioneiro dos estudos da terminologia. Ele considerava a terminologia como uma área de estudos interdisciplinar, passível de convergir com outras ciências, tais como a lógica, ciência da informação, a própria linguística, entre outras (Wüster, 1998, p. 26<sup>2</sup> *apud* Krieger, 2000, p. 211). A sua Teoria Geral da Terminologia (TGT) possui foco na criação de um sistema de comunicação especializada, uma vez que o meio científico, durante os anos 1930, carecia de um tratamento padronizado dos termos técnicos-científicos (Gomes, 2021).

Assim, “[...] a Teoria Geral da Terminologia assume um caráter metodológico, de natureza prescritiva e normalizadora em detrimento da

apreensão quanto aos modos de funcionamento dos léxicos terminológicos” (Krieger, 2000, p. 212-213). Isso porque Wüster defendia o uso dos termos técnico-científicos para normalizar a comunicação entre os acadêmicos da época. Para o teórico, os termos eram unívocos, ou seja, possuidores de somente um sentido.

Dessa forma, a TGT, além de uma abordagem metodológica, também se caracteriza pela abordagem onomasiológica. Na semântica, a abordagem onomasiológica leva em consideração o emissor, aquele que fala, enquanto a abordagem semasiológica, aponta para o receptor, aquele que ouve (Baldinger, 1966). Nesse sentido, Wüster visionava a criação dos sistemas de conceitos pela ótica de quem estava organizando os sistemas, não pela perspectiva de solucionar os possíveis problemas dos usuários.

Portanto, a teoria Wusteriana enfatiza a criação de sistemas de conceitos, de maneira a facilitar a fluidez da comunicação científica utilizando o arranjo sistemático, já que o sistema alfabético não era suficiente para o volume de conhecimento sendo produzido naquela época (Gomes, 2021). Para Krieger (2000), o léxico terminológico era tema principal da Teoria Geral da Terminologia, visando padronizar a linguagem especializada de determinada área do conhecimento, ou seja, os termos poderiam se referir a um conceito dentro de uma determinada área do conhecimento.

Em contrapartida à teoria de Wüster, Maria Teresa Cabré desenvolve a Teoria Comunicacional da Terminologia (TCT), a partir de 1993. Com o propósito de suprir as insuficiências da TGT, a qual ao decorrer do tempo passou a ser vista como uma teoria reducional da função da linguagem (Krieger, 2000), a Teoria Comunicacional da Terminologia busca contemplar a variação linguística, com uma visão comunicativa em que o foco é criar termos não somente para especialistas de uma área. Assim, a TCT “[...] considera que o uso de termos está sujeito ao princípio de adequação de acordo com o assunto, a função, os usuários, o tipo de texto, etc.” (Santamaria-Perez, 2009). Percebe-se, então, que enquanto a TGT defende a unidade terminológica como estática, a TCT afirma que outros fatores como condições pragmáticas e sociais irão influenciar na formação das unidades.

Os estudos de Cabré (1993, 2010<sup>3</sup>) apontam que os termos não são unívocos, pois um mesmo termo pode conter uma significação diferente em mais de uma área. Por exemplo, o termo “Reação” está presente na química, biologia e física, e terá um significado em cada uma delas.

Desse modo, a Teoria Comunicacional de Cabré se volta para a abordagem semasiológica, pois se preocupa com o receptor da mensagem, ou seja, o usuário, seja ele um especialista ou não. No cenário proposto por Cabré, é o usuário que vai utilizar o termo inserido em um instrumento de padronização (glossário, dicionário, tesouro, entre outros). Por esse aspecto, Almeida (2006) faz a relação teórico-prática da TCT, e afirma que a teoria oferece condições para a elaboração de produtos terminológicos, como os dicionários.

Nesse sentido, é visível o caráter descritivo da teoria de Cabré, que conversa bem mais com o tratamento descritivo da informação na Ciência da Informação, em comparação com a teoria de Wüster. Na TCT, a Terminologia, ciência que estuda os termos, tem o conceito como ponto de partida, possui característica descritiva de base linguística e faz parte de um campo multi e transdisciplinar (Santamaria-Perez, 2009). Considerando o tema discutido neste trabalho, é relevante fazer um paralelo com o campo da documentação, uma vez que o trabalho documental, ou análise conceitual, não poderia ser feito sem o uso de termos. Há, portanto, um intercâmbio entre a Terminologia e a Ciência da Informação, especificamente, com relação à indexação, através da qual se faz o emprego de termos a um determinado documento no processo de análise de assunto, como elabora Fujita (2003).

O trabalho de Cabré (2010) se volta para a criação de formas de representar o conhecimento especializado e de como o transmitir no domínio da Linguística. Já outros teóricos como Naves (2000), Neves (2004), Lancaster (2004), Fujita (2003; 2010), Redigolo (2014), abrem discussões a respeito da representação de assunto, também chamada de indexação, na Ciência da Informação.

A indexação, como um dos processos de representação da informação, consiste na leitura de partes específicas de um documento, com o objetivo de representá-lo por meio de termos.

Ao descrever os processos de indexação, Lancaster (2004, grifo nosso) relata as duas etapas essenciais: a **análise conceitual**, em que o indexador toma conhecimento do assunto de que trata o documento; e **tradução** dos termos retirados durante a análise conceitual, geralmente utilizando um vocabulário controlado. Em meio a isso, há a possibilidade de se atribuir termos por outras fontes além do próprio documento, prática que o autor denomina como indexação por atribuição.

Os termos atribuídos servem de ponto de acesso em uma base de dados para que o item seja localizado, recuperado e utilizado. Para isso, dois tipos de indexação podem ser realizados, a indexação **seletiva** e a **exaustiva**. A indexação seletiva procura empregar o menor número de termos a um documento, permitindo um acesso limitado a este, enquanto na indexação exaustiva se tem o maior número de termos identificados, permitindo maior possibilidade de acesso ao documento por um usuário.

Lancaster (2004, p. 9) sugere três perguntas norteadoras para a escolha de termos durante a indexação de documentos: “De que se trata? Por que foi incorporado no nosso acervo? Quais de seus aspectos serão de interesse para nossos usuários?”. O autor defende que um mesmo documento possa ser indexado de forma diferente, dependendo da clientela.

O conhecimento prévio do indexador fará a diferença no momento da indexação. Um indexador que tem mais familiaridade com determinado assunto, fará uma indexação minuciosa do material que está trabalhando. Nesse caso, Lima, Fujita e Redigolo (2021) afirmam que o contexto será um elemento fundamental para a escolha de conceitos durante a análise de assunto.

Autoras brasileiras como Naves (2000), Neves (2004), Fujita (2010), realizam estudos sobre a abordagem sociocognitiva na representação de assunto. Durante as análises feitas, as autoras constataram nos seus diferentes trabalhos que o tempo de experiência, conhecimento da área à qual o texto se refere, o modo de leitura, e o contexto profissional influenciam no processo de indexação. Ou seja, aspectos sociocognitivos podem influenciar na estratégia de identificação e seleção de conceitos para representar o assunto e na tomada de decisões sobre termos representativos e o porquê da sua escolha por parte do

indexador.

Além de livros e artigos, que geralmente é o objeto mais significativo para prática de indexação, é possível fazer a representação de diversos recursos informacionais, como CD, partituras, fotografias, vídeos e áudios. Logo, é necessário representar uma informação pelos dados que ela é composta, a fim de que seja recuperada pelo usuário final. Os metadados surgem com esse objetivo de descrever um objeto conforme suas características, de modo consistente e uniforme para a preservação dos objetos catalogados e sua posterior recuperação em meio virtual. Tipologias dos metadados poderão ser conferidas na seção 3 deste artigo.

A Ciência da Informação, segundo Borko (1968), preocupa-se com o vocabulário da área e a colaboração com outras sociedades. O autor destaca, também, a importância da manipulação da informação, o rápido avanço da tecnologia e a necessidade de atualização constante de conhecimentos, e o papel do cientista da informação na sociedade contemporânea.

Nesse sentido, é entendido que os cientistas, ou especialistas da informação têm papel fundamental na transferência de conhecimentos de um domínio, tendo as tecnologias da informação como instrumento para mediar informações. Hjørland e Albrechtsen (1995) trazem o conceito de análise de domínio para a Ciência da Informação, e explicam a quantidade de conhecimento que os especialistas precisam se apropriar para atuar em uma área.

Para Hjørland (2017, p. 4-5, tradução nossa), “[...] mediar o conhecimento do assunto requer um grau de conhecimento do assunto [...] as competências dos especialistas em informação são infra-estruturas de informação e recuperação da informação”. Para o autor, o estudo da terminologia específica e da recuperação da informação são dois tipos de qualificação da análise de domínio (Hjørland, 2017).

Franca e Silveira (2014) relatam que há diversas inconsistências na representação documental da literatura indígena em catálogos de bibliotecas. Isso pode ocorrer pela falta de orientação no código de catalogação *Anglo-American Cataloging Rules*, 2.ed. rev. (AACR2r) sobre ponto de acesso para

nomes indígenas. A incoerência no momento da representação de assunto, trará consequências na recuperação da informação pelos usuários, e nesse caso, pode até mesmo tornar invisível a produção literária de indígenas.

A representação documental de materiais produzidos por indígenas precisa ser abordada na literatura científica de Ciência da Informação, visto que a “[...] representação dos documentos de diferentes grupos sociais e étnicos é uma maneira de garantir a preservação da memória destes grupos [...]” (Franca; Silveira, 2014, p. 76).

Por outro lado, Santos (2008) não indicou o *Anglo-American Cataloging Rules*, 2.ed. rev. (AACR2r) para a catalogação de arquivos multimídia do acervo documental do setor de Linguística do Museu Paraense Emílio Goeldi, justamente por não abranger conteúdo da cultura indígena. Foi desenvolvido, então, um Código de Catalogação “[...] seguindo-se critérios estabelecidos, que deveriam ter representação da língua, data e origem do material” (Santos, 2008, p. 267).

Os critérios para representação a que Santos (2008) se refere ainda são utilizados pelo Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (ALIM/MPEG) para a nomenclatura dos arquivos, tais como: data de gravação, código ISO 639-3, sigla do pesquisador, sigla do falante nativo, termo representativo de conteúdo, numeração de gravação. Assim, percebe-se que por ser um acervo de conteúdo especializado, o tratamento da informação não será o mesmo que de um catálogo de uma biblioteca, por exemplo.

Ressalta-se que este artigo não propõe uma análise de domínio das línguas indígenas brasileiras, mas utiliza-se do princípio do estudo terminológico dentro de uma área e dos benefícios informacionais que a representação de assunto pode fornecer. É um estudo preliminar sobre a terminologia usada em arquivos digitais de um acervo específico. Dessa maneira, este estudo é circunstanciado pelo acervo de línguas indígenas como área de domínio terminológico.

Dito isso, as línguas indígenas da Amazônia constituem um domínio. Estima-se que haja em torno de 160 línguas indígenas faladas no Brasil, segundo o Instituto Socioambiental (2023). O Atlas das Línguas do Mundo em Perigo da

UNESCO (2010, p. 89) revela que ao menos um quinto das 150 línguas indígenas levantadas estavam em situação de ameaça, porém esse número pode ser ainda maior (Drude *et. al.*, 2023). Nesse cenário, o setor linguístico do Museu Paraense Emílio Goeldi tem colaborado para o processo de revitalização das línguas indígenas ameaçadas por meio da documentação linguística e a disponibilização desse material em um acervo digital.

Portanto, representar o assunto desses materiais oriundos de documentação linguística por meio de metadados faz parte do processo de documentação dessas línguas e pode contribuir para processos de revitalização das mesmas, uma vez que os materiais podem ser acessados pelas comunidades indígenas, em busca de aprendizagem e aprimoramento da sua língua nativa, ou por pesquisadores que têm interesse em desenvolver novas pesquisas em Linguística a partir dos dados armazenados no acervo digital.

### **3 Metadados: definição e tipologia**

Conforme muito comumente repetido na literatura de Ciência da Informação (CI), metadado significa dados sobre dados. Souza, Vendrusculo e Melo (2000, p. 93) acrescentam que metadado “[...] é a catalogação do dado ou descrição do recurso eletrônico”. Para Felipe Arakaki (2019, p. 62) os metadados são frequentemente usados em “[...] ações de representar, estruturar, gerenciar, preservar, usar e reusar informações”. Assim, considera-se metadado um dado que identifica e diferencia um recurso informacional a ser catalogado.

Com o avanço das tecnologias da informação e comunicação, os metadados têm mostrado sua importância no contexto informacional. Utilizado pela primeira vez no campo da Ciência da Computação, entre os anos 1960 e 1970, o termo metadado surgiu nos sistemas de banco de dados (Vellucci, 1998; Haynes, 2004). A Internet fez com que uma explosão de informações tivesse origem, o fato de que vários indivíduos pudessem criar e publicar conteúdos por meio de seus celulares tornaram os metadados uma fonte de análise e extração de informações sobre essas pessoas (Reis; Sá, 2020).

Dito isso, volta-se para o uso dos metadados na Ciência da Informação e no que eles representam para a descrição da informação, uma vez que a CI, não

somente possui os museus e bibliotecas como objeto de estudo, mas também a informação (Le Codiac, 2004) e todos os processos de uso e comunicação que a envolvem.

Para Alves (2010, p. 13), o tratamento descritivo da informação se refere à “individualização e caracterização de um item documentário”. A autora não compactua com o termo representação descritiva, por considerá-lo redundante já que toda representação é uma descrição. De todo modo, criar e usar metadados é uma atividade de tratamento descritivo, uma vez que se procura retirar de um documento/arquivo digital dados identificadores e representativos do mesmo, condizente com o contexto.

Para ampliar o conceito de metadados e suas aplicações, Ana Arakaki e Felipe Arakaki (2020) fizeram uma revisão de literatura sobre metadados, e constataram as seguintes tipologias:

**Quadro 1** - Tipologia de metadados segundo Ana Arakaki e Felipe Arakaki

<b>Tipologia de Metadados</b>	<b>Descrição</b>
Administrativos	Usados para gerenciar e administrar coleções e recursos informacionais, para auxiliar na tomada de decisão e na manutenção dos registros e recursos informacionais. Fornecem informações sobre a origem e a manutenção de um objeto;
Autenticação	São informações que possibilitam a identificação, integridade, legitimidade de um recurso informacional
Preservação	Estão relacionados com informações de preservação e conservação dos recursos informacionais
Proveniência	Estão relacionados às informações de procedência, fornecem dados sobre entidades, criação e modificações e seus relacionamentos;
Técnicos	Estão relacionados a como um sistema funciona, fornecem informações do sistema ou do recurso;
Meta-metadados (Metametadata)	Correspondem às informações sobre o registro criado, ou informações da criação de um conjunto de dados;
Descritivos	Identificam características identificadoras e os contextos intelectuais dos recursos de informação para fins de descoberta, identificação, seleção, aquisição, contexto e compreensão;
Direitos	Estão relacionados às informações sobre propriedade e direitos autorais;
Acesso e uso	São informações de como um recurso informacional foi acessado e utilizado, como restrições de circulação e acesso, registros de exposições, entre outros;
Estruturais	São informações de como um recurso informacional foi acessado e utilizado, como restrições de circulação e acesso, registros de exposições, entre outros;
Linguagens de marcação (Markup languages)	Integram metadados e sinalizações para outros recursos estruturais ou semânticos

Fonte: Arakaki e Arakaki (2020).

Percebe-se, então, a importância dos metadados para o gerenciamento da informação em diversos suportes. Dentre os tipos citados, ressalta-se os **metadados descritivos**, cujo uso seria adequado para a representação de arquivos digitais de um acervo multimídia, por exemplo. Esse tipo de metadado ordena dados característicos e específicos não somente com o foco na identificação do recurso informacional, mas também buscando a recuperação e aquisição do recurso informacional que o usuário precisa.

#### **4 Metodologia**

A partir da pesquisa descritiva, que consistiu no levantamento, por meio de documentos, como livros e artigos, sobre as teorias da terminologia, indexação e metadados, buscou-se fazer uma conexão teórico-prática referente à criação de metadados. Essa etapa da metodologia foi subdividida em três fases: (i) um estudo inicial do sistema de metadados utilizados no Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (ALIM/MPEG) para a identificação de fragilidades e inconsistências; (ii) a análise dos tipos de recursos informacionais presentes no acervo; (iii) o desenvolvimento e aplicação de uma política de catalogação/indexação para criação de novos metadados e padronização dos metadados já existentes e dos novos materiais a serem incluídos no acervo.

O programa Lameta (Hatton *et al.*, 2021) foi utilizado para a inserção dos metadados, durante o processo de catalogação do acervo linguístico. Trata-se de um programa de uso livre que organiza os metadados de coleções digitais, e tem como objetivo catalogar a documentação de línguas, assim como de outras formas de expressões culturais. O programa organiza as informações em três áreas principais: **Projeto, Sessão, Pessoas**. O uso do Lameta (Hatton *et al.*, 2021) se justifica por ser um dos programas mais adequados para a organização de coleções linguísticas, ultimamente. É um programa ainda em desenvolvimento, mas que simplifica o trabalho de inserir metadados sobre as gravações realizadas em campo.

A próxima etapa do estudo pode ser caracterizada como pesquisa-ação, na qual o pesquisador se torna participante do grupo estudado, criando um elo

entre a comunidade pesquisada e os pesquisadores que a estudam. Nessa etapa, foi selecionado um projeto de documentação linguística realizado por pesquisadores vinculados ao Museu Paraense Emílio Goeldi: o projeto intitulado *Documentação de duas línguas severamente ameaçadas Makurap e Wayoro (Brazil)*<sup>4</sup>, de Galucio, Nogueira e Costa (2023). Essas línguas são faladas, respectivamente, pelos povos indígenas Makurap e Ajuru da região do Rio Guaporé, em Rondônia. As gravações originais produzidas por esse projeto foram utilizadas para testar os procedimentos para a criação e padronização dos metadados no Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi e estabeleceu-se uma troca entre linguistas e tecnólogos da informação em todas as fases da execução do estudo.

Com o acesso a essas gravações, foi possível criar e revisar os metadados dos materiais e publicar em duas bases de dados: Endangered Languages Archive (ELAR) (ELAR, 2002) e Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (MPEG, 2023). Este acervo está sendo migrado para a plataforma DSpace, porém, por ainda se encontrar em fase de teste, o estudo fez uso das coleções indexadas na plataforma antiga.

Por isso, para avaliar a eficiência dos metadados criados e inseridos no acervo, foi feita uma comparação entre os metadados da língua Aikanã e as línguas Makurap e Wayoro para a discussão dos resultados, e num segundo momento, apresenta-se a análise da recuperação da informação no acervo digital, tendo duas coleções de línguas indígenas como exemplo a Makurap e a Wayoro.

O Endangered Languages Archive (ELAR, 2002) é a base de dados linguística do Endangered Languages Documentation Programme (ELDP), que publica as coleções dos documentadores e linguistas financiados pelo ELDP. O ELDP foi fundado em 2002, a partir do fundo filantrópico Arcadia Fund, com o objetivo de documentar e preservar idiomas ameaçados de extinção ao redor do mundo. Da mesma forma, o Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (MPEG, 2023) salvaguarda as coleções de 71 línguas indígenas da região Amazônica, que estão disponíveis para acesso na base de dados.

A fase final do estudo consistiu na elaboração de um manual atualizado,

visando descrever o passo a passo dos preenchimentos dos campos essenciais do Lameta de maneira que possa ser usado na representação da informação no ALIM, e otimizar a criação de metadados dos arquivos multimídia utilizando as teorias da terminologia e indexação, assim como a otimização do processo de recuperação da informação. Dessa forma, dividem-se as etapas metodológicas em sete momentos:

**Quadro 2 - Etapa metodológica**

<b>Etapas</b>	<b>Operalização</b>
Extração de termos do material multimídia	Extraír termos representativos ao conteúdo do vídeo/áudio/foto e conceder a nomenclatura ao arquivo conforme o modelo padrão do setor.
Criação dos metadados	Extraír dados referentes a vídeos, áudios, imagens, e outros recursos informacionais que serão armazenados no acervo.
Revisão do material multimídia	Examinar o material multimídia (vídeos, áudios, fotos, entre outros) para a extração dos metadados. Nesta etapa, os arquivos podem ser convertidos, recortados, concatenados.
Preenchimento das planilhas com os metadados	Preencher as duas planilhas do Lameta: “People” e “Session” com os dados das pessoas envolvidas no projeto e os metadados dos arquivos multimídia, respectivamente.
Importação das planilhas no Lameta	Importar as duas planilhas do Lameta: “People” e “Session” preenchidas com os metadados para a interface do programa.
Revisão dos metadados criados	Avaliar os metadados criados, propor alterações adequadas se necessário para a melhoria na representação da informação.
Exportação dos IMDI e Opex Files	Exportar o Isle Meta Data Initiative – IMDI para a interoperabilidade entre diferentes ferramentas de leitura digital. Exportar o material multimídia para a base de dados Endangered Languages Archive e Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi.

Fonte: Elaborado pelas autoras.

## 5 Resultados e discussão

Nesta seção, apresentam-se os resultados comparativos entre o padrão de metadados usados inicialmente no Acervo de Línguas Indígenas do Museu Goeldi e o padrão definido a partir da pesquisa desenvolvida com o projeto *Documentação de duas línguas severamente ameaçadas Makurap e Wayoro (Brazil)* (Galucio; Nogueira; Costa, 2023) conforme citado na metodologia, no qual se trabalhou as coleções das línguas indígenas Makurap (2023) e Wayoro (2023).

A comparação foi feita com os metadados descritivos mais significativos para a representação das informações contidas nos arquivos, como: título, data, descrição, palavras-chave e gênero.

Em segundo momento, faz-se uma análise da recuperação da informação no ALIM/MPEG, a partir da indexação dos arquivos multimídia com os novos metadados. Considera-se fundamental essa etapa, visto que “[...] a indexação e recuperação da informação são processos indissociáveis, pois, o primeiro influencia diretamente os efeitos do segundo [...]”, conforme Garcia e Redigolo (2020, p. 6).

### **5.1 Em relação à criação e padronização de metadados**

Dentre as línguas indígenas catalogadas no acervo do MPEG, a coleção da língua Aikanã (2023) foi escolhida para servir de exemplo para esta seção, já que a coleção apresenta uma quantidade significativa de arquivos coletados, em comparação às outras línguas do acervo, possibilitando material suficiente para a comparação dos metadados.

O quadro 3 apresenta os metadados mais relevantes para identificação dos arquivos do acervo multimídia. As três colunas se referem às línguas indígenas Aikanã, Makurap e Wayoro e as respectivas descrições dos metadados utilizados. Os elementos estão em inglês no acervo digital do MPEG, porém, para melhor visualização, compreensão e redação do artigo, optou-se por traduzir os elementos para o Português.

**Quadro 3** - Comparação entre metadados das línguas indígenas Aikanã, Makurap e Wayoro

Elementos	Aikanã	Makurap	Wayoro
<b>Título</b>	Palavras verbais com MA	Elicitação lexical sobre macaxeira	Elicitação de palavras e frases do dia a dia
<b>Data</b>	2009-09-16	2023-01-17	2023-08-08
<b>Descrição</b>	Esta sessão <sup>5</sup> abrange apenas uma parte dos registros em áudio, vídeo e foto sobre a língua e a cultura do povo Aikanã. Que tem por finalidade documentar a língua e a cultura do povo. Em específico esta sessão concerne na elicitação de uma lista de palavras sobre verbos futuros na língua Aikanã.	Esta sessão concerne na elicitação lexical sobre partes da planta da mandioca e alimentos feitos com mandioca. A pesquisadora (Vilacy Galucio) faz as perguntas em português e a D. Isaura Makurap e a D. Marina Makurap respondem na língua Makurap. A D. Isaura mencionou alguns alimentos feitos com mandioca e explicou como prepará-los, como o mayĩ kat (moqueca com mandioca e gongo).	Esta sessão é sobre a elicitação de palavras e frases do dia a dia, como alimentação e pintura corporal, com Paulina Macurap e Maria Ajuru.
<b>Gênero</b>	Estímulo (stimuli)	Elicitação	Elicitação
<b>Palavras-chave / Assunto</b>	Lista de palavras; Pronomes verbais; Elicitação	Cassava / Macaxeira, Traditional foods / Comidas tradicionais	Community's daily life / Dia a dia da comunidade, Fishes / Peixes, Ways of cooking / Formas de cozinhar, Fruits / Frutas, Body painting / Pintura Corporal

Fonte: Elaborado pelas autoras.

No que se refere ao elemento ‘Título’, Aikanã apresenta um título correspondente ao conteúdo do arquivo, sobre uma elicitação de verbos com um falante nativo (MA) da língua. Os títulos das coleções Makurap e Wayoro apresentam maior similaridade um com o outro, já que fazem parte do mesmo projeto de documentação, e ainda assim, também correspondem ao conteúdo dos arquivos com mais clareza. Ter o tema principal do arquivo representado no título pode gerar uma forma de recuperação relevante para o usuário.

Sobre o elemento ‘descrição’, a coleção de Aikanã (2023) apresenta um parágrafo inicial aparentemente padrão, já que em várias outras sessões o mesmo parágrafo se repete antes de abordar especificamente o conteúdo do arquivo em questão em um segundo parágrafo. Por outro lado, as descrições das línguas Makurap (2023) e Wayoro (2023) são redigidas objetivamente,

provendo informações sobre as ações dos envolvidos nas gravações e o assunto abordado durante as mesmas. É importante que este campo seja preenchido com informações relevantes sobre o conteúdo, já que se destina à representação temática do recurso informacional.

Em relação aos elementos ‘gênero’ e ‘palavras-chave / assunto’, Aikanã foi catalogada como “estímulo”. Isso se deu devido à limitação do programa utilizado na época da catalogação no ALIM, o qual não possuía a opção de escolher elicitación como uma das opções no campo gênero. Já Makurap e Wayoro puderam ser catalogadas com o gênero “elicitación” porque o programa usado para as duas coleções, Lameta, permitiu que esse gênero pudesse ser selecionado para sessões específicas.

Ainda, Makurap e Wayoro formaram coleções bilíngues por conta das especificações da instituição fomentadora do projeto, por isso as palavras-chave estão em dois idiomas. Pode ser observado um nível maior de especificidade nessas duas coleções com as palavras-chave “comidas tradicionais”, “dia a dia da comunidade”, “formas de cozinhar” e “pintura corporal”. Fujita (2020) discute o uso da linguagem controlada e a linguagem natural nas palavras-chave de artigos científicos. Para a autora, é essencial que haja uma coexistência entre esses dois tipos de linguagens na representação de assunto, visto que a linguagem natural é mais familiar para o usuário.

Sugere-se, então, que os termos referentes ao assunto do arquivo possam responder duas das três perguntas norteadoras listadas por Lancaster (2004): Do que se trata? Quais de seus aspectos serão de interesse para nossos usuários? Alguns termos já parecem responder essas indagações, como “dia a dia da comunidade” (sobre o que trata o arquivo) e “macaxeira” (relevante para o usuário).

Da mesma maneira, sugere-se que para a redação do campo ‘descrição’ se faça as duas perguntas norteadoras de Lancaster (2004), explicando o conteúdo dos arquivos, a relevância dele para o acervo, os detalhes do projeto e dos motivos que levaram os pesquisadores a documentar aquele tipo de conteúdo. Além disso, o número de termos extraídos dos arquivos difere em cada sessão, por isso não é possível dizer, certamente, se a indexação feita foi

seletiva ou exaustiva.

As palavras-chave de Makurap e Wayoro são mais próximas à linguagem do usuário do que à linguagem controlada dos vocabulários especializados. Aponta-se, também, que o acervo de línguas indígenas é de caráter especializado e não se conhece algum tesouro sobre línguas indígenas que poderia se aplicar ao contexto, de forma que os termos para a representação de assunto foram extraídos diretamente dos conteúdos dos arquivos e permaneceram na linguagem natural.

Percebe-se que os novos metadados criados para as coleções de Makurap e Wayoro demonstram uma representação coesa com os assuntos dos arquivos de mídia, de modo que possibilita a identificação dos temas e maior visibilidade dos arquivos. Os elementos Título, Descrição e Palavras-chave formam um conjunto de elementos representativos da temática dos arquivos de mídia. Zeng e Qin (2016) relatam que os objetos de museus não carregam em si informação descritiva, sendo necessário que os indexadores criem esta informação. Da mesma maneira, os arquivos de mídia precisam que os metadados sejam elaborados pelos pesquisadores em campo, durante ou depois das gravações, para que o arquivo seja catalogado e indexado posteriormente.

Os metadados descritivos controlam a dispersão informacional, e mantêm o acervo mais organizado e de fácil acesso para usuários. Tendo os usuários em mente, a representação de assunto do ALIM se interliga com a abordagem semasiológica, que se preocupa com o receptor da mensagem, tal como definido na Teoria Comunicacional da Terminologia (TCT) de Cabré (1993). Percebe-se a flexibilização dos termos no acervo, ao invés da característica unívoca dos termos técnicos-científicos sugerido por Wüster (2010).

## **5.2 Em relação à avaliação da recuperação da informação**

O Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (MPEG, 2023) possui 71 coleções indexadas virtualmente até o momento de elaboração deste trabalho. Esta seção examina os metadados descritos, especificamente, nas coleções das línguas Makurap (2023) e Wayoro (2023) incluindo tanto as

sessões já existentes e organizadas de acordo com o padrão antigo de nomenclatura e demais metadados, quanto às novas sessões organizadas de acordo com o padrão de indexação e organização de metadados definidos na pesquisa aqui apresentada.

A coleção de Makurap (2023) possuía nove sessões indexadas no acervo, após a inclusão de 30 novas sessões com os metadados propostos no presente estudo, a coleção conta com um total de 39 sessões, considerando todas as categorias de assunto. A coleção de Wayoro (2023), possuía 71 sessões indexadas no acervo. Acrescentaram-se 18 novas sessões, totalizando 89 sessões, considerando todas as categorias de assunto.

O objetivo foi verificar a influência dos metadados na recuperação da informação. Os descritores usados na busca foram escolhidos visando estabelecer uma comparação entre a recuperação nos dois acervos das línguas indígenas Makurap e Wayoro de assuntos que foram incluídos na base de dados em ambas coleções. Para isso, então foram usados os descritores *elicitação*, descritor que se refere ao tipo de conteúdo, e *pamonha*, descritor que se refere a um assunto ou tema específico. Como forma de controle, antes de realizar a busca com os descritores, foi feito um levantamento independente de quantas sessões indexadas havia em cada língua sobre *elicitação* e *pamonha*. Esse número é apresentado na tabela 1.

**Tabela 1** - Sessões recuperadas com o descritor *elicitação*

Língua Indígena	Nº de sessão indexadas inicialmente no acervo sobre <i>elicitação</i>	Nº de sessões acrescentadas no acervo sobre <i>elicitação</i>	Nº total de sessões sobre <i>elicitação</i>	Nº de sessões recuperadas com o descritor <i>elicitação</i>	Campos
Makurap	7	11	18	16	Categoria, Chaves, Descrição, Subgênero, Título,
Wayoro	33	1	34	35	Categoria, Chaves, Descrição, Subgênero,

Fonte: Elaborado pelas autoras.

O resultado da busca mostra que a recuperação da informação foi

adequada, mas pode haver melhorias a partir da revisão dos metadados. De 18 sessões sobre eliciações em Makurap, 16 foram recuperadas. As outras duas sessões faltantes não apareceram na busca porque não mencionaram a palavra “elicitação” em nenhum campo, já que se tratavam, principalmente, de sessões sobre a gramática da língua. Mas ainda assim, se encaixam na categoria de eliciação-gramática-fonologia do ALIM.

Uma possibilidade de melhoria na criação dos metadados para otimizar os resultados da busca é garantir que pelo menos um dos campos da sessão faça referência à categoria onde a sessão está indexada. Ou seja, a indexação exaustiva explicada por Lancaster (2004) pode ser a solução para este caso, uma vez que aumenta a possibilidade de acesso a partir dos vários termos atribuídos aos arquivos e servem como ponto de acesso.

Das sessões sobre eliciações em Wayoro, 35 foram recuperadas, ou seja, uma sessão a mais do que o esperado. Esse resultado ocorreu porque uma sessão sobre manufatura está registrada, no campo Categoria padrão (*Task*) como “elicitação-léxico”, ao invés de estar registrada como uma atividade cotidiana, ou seja, um erro na definição dos metadados resultou em um falso positivo, quando da realização da busca.

**Tabela 2** - Sessões recuperadas com o descritor *pamonha*

Língua Indígena	Nº de sessão indexadas inicialmente no acervo sobre <i>pamonha</i>	Nº de sessões acrescentadas no acervo sobre <i>pamonha</i>	Nº de sessões recuperadas com o descritor <i>pamonha</i>	Campos identificados
Makurap	0	1	1	Descrição
Wayoro	0	2	2	Nome, Título, Descrição, Assunto

Fonte: Elaborado pelas autoras.

Após a inclusão de sessões com novos metadados, o acervo de Wayoro continua, ao todo, duas sessões sobre o assunto *pamonha*. Ambas foram recuperadas quando o descritor *pamonha* foi usado para a busca. Já a coleção de Makurap citava *pamonha* em uma sessão de eliciação lexical, que também foi recuperada. O resultado teve, portanto, alto nível de precisão e revocação, ou seja, quando o sistema é capaz de evitar documentos inúteis e recuperar

documentos úteis aos usuários, respectivamente (Lancaster, 2004).

Cabré (1993) diz que uma forma de representar a realidade especializada é por meio da terminologia. Levando em consideração o contexto dos povos indígenas, o milho, que dá origem à pamonha, é um alimento que pode ser consumido cozido, assado ou servido como bebida fermentada, por isso é importante o registro sobre tudo o que envolve esse cereal, e a representação dessa realidade pela terminologia.

Portanto, verificou-se que a criação de metadados para a indexação dos arquivos multimídia pode ter grande influência na recuperação da informação pelos usuários do acervo linguístico, uma vez que os processos estão relacionados. Por isso é importante que os dados sejam descritos de forma coerente e objetiva em relação ao conteúdo dos arquivos, e nos devidos campos. Por exemplo, uma estratégia é utilizar a linguagem natural, que é mais próxima dos usuários, nos campos de descrição e assunto, e sempre verificar se as categorias de assunto foram registradas corretamente em uma sessão.

## **6 Considerações finais**

O estudo abordou as teorias da Terminologia e Indexação para a criação de metadados, e avaliou de maneira comparativa a descrição dos metadados atuais do Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi (ALIM/MPEG) com os metadados criados no projeto *Documentation of the severely endangered languages Makurap and Wayoro (Brazil): material and non-material traditional culture, and its associated knowledge* (Galucio; Nogueira; Costa, 2023), que foi utilizado como base de teste para a proposta deste estudo para a padronização dos termos usados nos metadados.

Os metadados, devidamente extraídos dos conteúdos, otimizam os processos de representação e recuperação da informação. Os estudos de Cabré (1993, 2010) e Lancaster (2004) se mostraram mais favoráveis ao tipo de linguagem utilizada nos termos representativos do ALIM, ou seja, a linguagem natural.

Foi realizada também uma análise sobre a recuperação da informação no ALIM com as coleções de Makurap e Wayoro, considerando os arquivos que

foram catalogados no Lameta, utilizando a leitura técnica de Lancaster (2004) para a extração de termos a partir dos próprios arquivos, na linguagem natural. Ressalta-se que há uma nova base de dados em elaboração, e o processo de criação dos metadados e migração dos arquivos está em andamento e em fase de testes, logo, os metadados proposto estão passíveis de revisão e melhoria para compor a nova base de dados do setor linguístico do Museu Paraense Emílio Goeldi.

A partir do objetivo proposto, constatou-se que a criação de metadados condizentes com o conteúdo dos arquivos de mídia possibilita a identificação prévia do que trata o recurso informacional para o usuário, assim como fornece melhor organização do acervo. Por conta do uso comum das tecnologias da informação, os metadados são elementos que fazem parte da indexação, processo que pode ser influenciado pelo contexto e pela tomada de decisões do indexador.

Portanto, o processo de criação de metadados pode ser padronizado por meio de uma política de indexação ou de um manual de criação de metadados, com o objetivo de formar um acervo harmônico, com dados descritivos que ressaltem o conteúdo dos arquivos catalogados, deem visibilidade aos arquivos e evitem sentenças repetitivas nas descrições das sessões. O manual para a criação de metadados pode, inclusive, orientar sobre a indexação seletiva ou exaustiva, o número de termos a ser extraído para cada sessão, considerando o conteúdo do arquivo e o interesse do usuário. Essa política de indexação e padronização dos metadados fornece bases satisfatórias para a recuperação da informação que satisfaça as necessidades informacionais dos usuários.

### **Financiamento**

Estudo financiado pela bolsa do Programa de Capacitação Institucional no Museu Paraense Emílio Goeldi (PCI/MPEG) para a primeira autora e bolsa de Produtividade em Pesquisa do Conselho Nacional de Pesquisa e Desenvolvimento (PQ/CNPq) para a segunda autora.

## Referências

- AIKANÃ. *In*: MUSEU PARAENSE EMÍLIO GOELDI (ALIM). Acervo de Línguas Indígenas, 2023. Coleção Aikanã.
- ALMEIDA, Gladis Maria de Barcellos. A Teoria Comunicativa da Terminologia e a sua prática. **Alfa: Revista de Linguística**, São Paulo, v. 50, n. 2, p. 85-101, 2006.
- ALVES, Rachel Cristina Vesú. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.
- ARAKAKI, Felipe Augusto. **Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV**. 2019. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2019.
- ARAKAKI, Ana Carolina Simionato; ARAKAKI, Felipe Augusto. Dados e metadados: conceitos e relações: concepts and relationships. **Ciência da Informação**, Brasília, v. 49, n. 3, p. 34-45, 2020. Disponível em: <https://doi.org/10.18225/ci.inf.v49i3.5504>. Acesso em: 14 nov. 2023.
- BALDINGER, Kurt. Semasiologia e onomasiologia. **Alfa: Revista de Linguística**, São Paulo, v. 9, p. 7-36, 1966.
- BORKO, Harold. Information science: what is it? **American Documentation**, United States, v. 19, n. 1, p. 3-5, 1968. (Tradução Livre).
- CABRÉ, Maria Teresa. **La terminología: teoría, metodología, aplicaciones**. Barcelona: Antártida, 1993.
- CABRÉ, Maria Teresa. **La terminología: representación y comunicación**. Barcelona: Universitat Pompeu Fabra, 2010.
- DRUDE, Sebastian; BIRCHALL, Joshua; GALUCIO, Ana Vilay; MOORE, Denny; VAN DER VOORT, Hein. Endangered Languages in Brazil in 2021. *In*: Eda Derhemi; Christopher Moseley (org.). **Endangered languages in the 21st Century**. London: Routledge, 2023, p. 21-40.
- ENDANGERED LANGUAGES ARCHIVE (ELAR). Berlin-Brandenburg: Endangered Languages Documentation Project, 2002.
- FRANCA, Aline; SILVEIRA, Naira Christofolletti. A Representação descritiva e a produção literária indígena brasileira. **Transinformação**, Campinas, v. 26, n. 1, p. 67-76, 2014.

FUJITA, Mariângela Spotti Lopes. Linguagem natural ou linguagem controlada? Influência da palavra-chave na representação para indexação e recuperação de informações. **Informação & Sociedade: Estudos**, João Pessoa, v. 30, n. 4, p. 1-29, 2020. Disponível em: <https://doi.org/10.22478/ufpb.1809-4783.2020v30n4.57041> Acesso em: 31 maio 2024.

FUJITA, Mariângela Spotti Lopes. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**; Campinas, v. 1, n. 1, p. 60-90, 2003. Disponível em: <https://doi.org/10.20396/rdbci.v1i1.2089>. Acesso em: 28 nov. 2023.

FUJITA, Mariângela Spotti Lopes. O contexto profissional do indexador no ensino de indexação. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, v. 15, n. 30, p. 91-104, 2010. Disponível em: <https://doi.org/10.5007/1518-2924.2010v15n30p91>. Acesso em: 28 nov. 2023.

GALUCIO, Ana Vilacy; NOGUEIRA, Antônia Fernanda; COSTA, Carla Daniele. [Makurap: Documentation of language and culture. Wayoro: Documentation of language and culture.] [S. l.], 2023. Disponível em: <https://elararchive.org/collections/>. Acesso em: 31 maio 2024.

GARCIA, Valdenice César; REDIGOLO, Franciele Marques. Indexação e recuperação da informação. **Tendências da Pesquisa Brasileira em Ciência da Informação**, São Paulo, v. 13 n. 1, p. 1-20, 2020.

GOMES, Hagar Espanha. Terminologia e estrutura conceitual. **PontodeAcesso**, Salvador, v. 15, n. 3, p. 184-203, 2021. Disponível em: <https://doi.org/10.9771/rpa.v15i3.47464>. Acesso em: 14 nov. 2023.

HATTON, John; HOLTON, Gary; SEYFEDDINIPUR, Mandana; THIEBERGER, Nick. **Lameta** [software]. [S. l.], 2021. Disponível em: <https://github.com/onset/laMETA/releases>. Acesso em: 14 nov. 2023.

HAYNES, David. **Metadata for information management and retrieval**. Cambridge: Facet Publishing, 2004. v. 1.

HJØRLAND; Birger; ALBRECHTSEN, Hanne. Toward a new horizon in information science: domain analysis. **Journal of the American Society for Information Science**, New Jersey, v. 46, n. 6, p. 400-425, 1995. Disponível em: [https://doi.org/10.1002/\(SICI\)1097-4571\(199507\)46:6<400:AID-ASI2>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400:AID-ASI2>3.0.CO;2-Y). Acesso em: 7 maio de 2024.

HJØRLAND; Birger. Domain Analysis. **Knowledge Organization**, Toronto, v. 44, n. 6, p. 436-464, 2017.

INSTITUTO SOCIOAMBIENTAL. **Povos Indígenas do Brasil - Línguas**, Brasil, 21 jun. 2023.

KRIEGER, Maria da Graça. Terminologia revisitada. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, São Paulo, v. 16, n. 2, p. 209-228, 2000. Disponível em: <https://doi.org/10.1590/S0102-44502000000200001>. Acesso em: 28 nov. 2023.

LANCASTER, Frederick Wilfrid. **Indexação e resumos: teoria e prática**. 2. ed. Brasília: Briquet de Lemos, 2004.

LE CODIAC, Yves-François. **A Ciência da Informação**. 2. ed. Brasília: Briquet de Lemos, 2004.

LIMA, Gercina Ângela Borém de Oliveira; FUJITA, Mariângela Spotti Lopes; REDIGOLO, Franciele Marques. A importância do contexto para a indexação. **PontodeAcesso**, Salvador, v. 15 n. 3, p. 283-302, 2021. Disponível em: <https://doi.org/10.9771/rpa.v15i3.47469>. Acesso em: 28 nov. 2023.

MAKURAP. *In*: MUSEU PARAENSE EMÍLIO GOELDI (ALIM). **Acervo de Línguas Indígenas**, 2023. Coleção Makurap.

MUSEU PARAENSE EMÍLIO GOELDI (MPEG). **Acervo de Línguas Indígenas**. Belém, 2023. Coleção: Acervo Linguístico.

NAVES, Madalena Martins Lopes. **Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores**. 2000. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2000.

NEVES, Dulce Amelia de Brito. **Aspectos metacognitivos na leitura do indexador**. 2004. Tese (Doutorado em Ciência da Informação) Escola de ciência da informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

REDIGOLO, Franciele Marques. O processo de análise de assunto na catalogação em bibliotecas universitárias para proposta de normalização: estudo de observação com protocolo verbal. **Brazilian Journal of Information Science: Research Trends**, Marília, v. 8, n. ½, p. 1-24, 2014. Disponível em: <https://doi.org/10.36311/1981-1640.2014.v8n1e2.07.p143>. Acesso em: 14 nov. 2023.

REIS, Luiz Carlos Rezende; SÁ, Maria Irene da Fonseca e. Big data: um novo campo de atuação para bibliotecários. **Prisma.com**, Porto, n. 41, p. 231-250, 2020.

RILEY, Jenn. **Understanding metadata**. Washington: National Information Standards Organization, 2017.v. 23.

SANTAMARÍA-PÉREZ, Isabel. **La Terminología**: definición, funciones y aplicaciones. Alicante: Universidad de Alicante, 2009. Disponível em: <http://hdl.handle.net/10045/12770>. Acesso em: 28 nov. 2023.

SANTOS, Ellison Cleyton Barbosa. Sistema de informação do acervo de línguas indígenas (SIALI): uma proposta de organização de acervo multimídia. **MOARA - Revista Eletrônica do Programa de Pós-Graduação em Letras**, Belém, n. 30, p. 255-269, jul./dez., 2008. Disponível em: <http://dx.doi.org/10.18542/moara.v2i30.3385>. Acesso em: 7 maio 2024.

SOUZA, Marcia Isabel Fugisawa; VENDRUSCULO, Laurimar Gonçalves; MELO, Geane Cristina. Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core. **Ciência da Informação**, Brasília, v. 29, n. 1, 2000. Disponível em: <https://doi.org/10.18225/ci.inf.v29i1.903>. Acesso em: 28 nov. 2023.

UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION (UNESCO). **Atlas das Línguas do Mundo em Perigo**. 3. ed. Paris: UNESCO, 2010.

VELLUCCI, Sherry L. Metadata. **Annual review of information science and technology (ARIST)**, United States, v. 33, p. 187-222, 1998.

ZENG, Marcia Lei; QIN, Jian. **Metadata**. 2nd. ed. Chicago: Neal-Schuman, 2016.

WAYORO. *In*: MUSEU PARAENSE EMÍLIO GOELDI (ALIM). Acervo de Línguas Indígenas, 2023. Coleção Wayoro.

WÜSTER, Eugen. **Introducción a la Teoría general terminológica y a la lexicografía terminológica**. Barcelona: IULA, 2010.

## **Descriptive metadata in digital collections: a case study in the Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi**

**Abstract:** With recent technical and technological advances in the field of information, there has been a proliferation of digital files, so it is important to control information in digital databases. The use of a standardized metadata system can be a solution to the problem of information dispersion since the metadata standard manages and organizes data according to its functionality. This article aims to analyze the strategies developed to optimize the cataloging and information retrieval process for the indigenous language collection at the Museu Paraense Emílio Goeldi. This study seeks to apply the principles of the General Theory of Terminology, the Communicational Theory of Terminology,

and the Theory of Indexing, linking them to the areas of Linguistics and Information Science. The methodology consisted of an analysis of the types of information resources present in the collection of indigenous languages at the Museu Paraense Emílio Goeldi and the development and application of a cataloging/indexing policy for the creation of metadata and the standardization of existing metadata and new materials to be included in this collection. The initial results showed that the creation of metadata consistent with the content of the media files makes it possible to identify the information resource in advance. The process of creating descriptive metadata can be standardized through an indexing policy or a manual on metadata creation, in order to provide a satisfactory basis for users' information needs.

**Keywords:** metadata; terminology; indexing; information retrieval

Recebido: 21/12/2023

Aceito: 22/05/2024

**Declaração de autoria:**

**Concepção e elaboração do estudo:** Stela Andrade Vasconcelos, Ana Vilacy Galucio

**Coleta de dados:** Stela Andrade Vasconcelos

**Análise e interpretação de dados:** Stela Andrade Vasconcelos, Ana Vilacy Galucio

**Redação:** Stela Andrade Vasconcelos, Ana Vilacy Galucio

**Revisão crítica do manuscrito:** Ana Vilacy Galucio

**Como citar**

VASCONCELOS, Stela Andrade; GALUCIO, Ana Vilacy. Metadados descritivos em acervos digitais: um estudo de caso no Acervo de Línguas Indígenas do Museu Paraense Emílio Goeldi. **Em Questão**, Porto Alegre, v. 30, e-137619, 2024. DOI: <https://doi.org/10.1590/1808-5245.30.137619>

**Parecer(es) aberto(s):**

<https://doi.org/10.1590/1808-5245.30.137619.A>

<https://doi.org/10.1590/1808-5245.30.137619.B>



---

<sup>1</sup> Termo no original: *Markup languages*.

<sup>2</sup> WÜSTER, E. Introducción a la teoría general de la terminología y a la lexicografía terminológica. Barcelona: Universitat Pompeu Fabra, 1998. *Apud* Krieger (2000).

<sup>3</sup> A primeira versão da obra é do ano de 1999, mas a autora deste artigo teve acesso a versão de 2010.

<sup>4</sup> Esse projeto é financiado pelo Endangered Language Documentation Program (ELDP) e seu título original em inglês é *Documentation of the severely endangered languages Makurap and Wayoro (Brazil): material and non-material traditional culture, and its associated knowledge*.

<sup>5</sup> No acervo, a palavra “sessão” foi redigida como “seção”. A troca de palavras no artigo foi feita para fins ortográficos.