XIX encontro nacional de pesquisa em ciência da informação
ENANCIB
// SUJEITO INFORMACIONAL E AS PERSPECTIVAS ATUAIS EM CIÊNCIA DA INFORMAÇÃO. //
22-26 OUTUBRO 2018
LONDRINA/PR
Universidade Estadual de Londrina

XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018

GT-8 – Informação e Tecnologia

MUSEU DO ÍNDIO: ESTUDO DE CASO DO PROCESSO DE MIGRAÇÃO E ABERTURA DOS DADOS LIGADOS SEMÂNTICOS DO ACERVO MUSEOLÓGICO COM O SOFTWARE LIVRE TAINACAN

Dalton Lopes Martins (Universidade de Brasília – UNB)

Danielle do Carmo (Universidade Federal de Goiás - UFG)

Leonardo Barbosa Germani (Universidade Federal de Goiás - UFG)

MUSEU DO ÍNDIO: CASE STUDY OF THE MIGRATION PROCESS AND OPENING OF THE SEMANTIC CONNECTED DATA OF THE MUSEOLOGICAL COLLECTIONS WITH THE FREE SOFTWARE TAINACAN

Modalidade da Apresentação: Comunicação Oral

Resumo: O trabalho apresenta um estudo de caso de migração da documentação do acervo museológico do Museu do Índio para o software livre Tainacan. O Museu do Índio é uma importante instituição pública federal brasileira ligada a Fundação Nacional do Índio criada em 1953 e possuí como missão a preservação e promoção do patrimônio cultural dos povos indígenas. Desta forma, o presente trabalho tem como objetivos apresentar e descrever os resultados obtidos a partir de 07 passos metodológicos que envolvem desde a análise técnica dos acervos, a coleta de informações, tratamento da informação até a migração da solução atual. Os resultados evidenciam o trabalho quantitativo de tratamento da informação, reduzindo os problemas técnicos sintáticos, normalização, reconciliação e desambiguação, chegando a reduzir, em média, em torno de 20% a variabilidade dos valores presentes nos metadados utilizados. Ressalta-se também os resultados obtidos na etapa de enriquecimento da informação por meio de recursos semânticos de reconciliação dos nomes das etnias indígenas com a plataforma Wikidata, chegando ao resultado em torno de 90% dos nomes identificados, conectando o acervo com outros recursos da Wikipédia e descrições multilíngues das etnias. Por fim, apresenta-se os resultados obtidos na interface gráfica e os ganhos em potencial para os usuários interessados no acervo do museu.

Palavras-Chave: Tainacan; Wikidata; Museu do Índio; Acervos em rede, Repositório digital.

Abstract: The paper presents a case study on the documentation migration of museological collections from the *Museu do Índio* to the free software Tainacan. The *Museu do Índio* is an important Brazilian federal public institution linked to the *Fundação Nacional do Índio*, that was created in 1953 and has as its mission the preservation and promotion of the cultural heritage of indigenous peoples. In this way, the present work aims to present and describe the results obtained from 07 methodological steps that involve from the technical analysis of the collections, information collection, information processing to the migration of the current solution. The results evidenced the quantitative work of information processing, reducing syntactic technical problems, normalization, reconciliation and disambiguation, reducing the variability of the values in the metadata used, on average, by around 20%. It is also worth mentioning the results obtained in the stage of information enrichment by means of semantic resources to reconcile the names of the indigenous ethnic groups with the Wikidata plataform, reaching the result of around 90% of the identified names, connecting the collection with other resources of the Wikipedia and multilingual descriptions of ethnicities. Finally, we present the results obtained in the graphic interface and the potential gains for users interested in the collection of the museum.

Keywords: Tainacan, Wikidata, Museu do Índio, Digital Collections, Digital Repository.

1 INTRODUÇÃO

O Museu do Índio é uma importante instituição pública federal brasileira ligada a Fundação Nacional do Índio (FUNAI). Criado em 1953, o museu tem como missão

> Preservar e promover o patrimônio cultural dos povos indígenas por meio de pesquisa, documentação, divulgação e diversas ações de fortalecimento de suas línguas, culturas e acervos, prioritariamente aqueles em situação de vulnerabilidade (MUSEU DO ÍNDIO, 2018).

O museu possui um dos mais importantes acervos relacionados à cultura indígena no país

> Tem sob sua guarda documentos relativos à maioria das sociedades indígenas contemporâneas, constituídos de 15 mil 840 peças etnográficas e 15 mil 121 publicações nacionais e estrangeiras, especializadas em etnologia e áreas afins. Seus diversos serviços são responsáveis pelo tratamento técnico de 76.821 registros audiovisuais e 833.221 documentos textuais de valor histórico e contemporâneo (MUSEU DO ÍNDIO, 2018).

Trata-se, portanto, de uma instituição de grande valor cultural para um país como o Brasil, apresentando não apenas uma grande riqueza de seu acervo, como os números acima denotam, mas também uma variedade informacional por se tratar de uma instituição que congrega os papéis de arquivo, museu e biblioteca. Dessa forma, convive cotidianamente com grande complexidade informacional, envolvendo diferentes tecnologias de organização do acervo, tais como padrões de metadados, regras de catalogação, linguagens documentárias, bases de dados, técnicas de digitalização de acervo, entre outros. Gerenciar tal complexidade impõe desafios das mais variadas ordens, sobretudo em relação às práticas

de gestão de acervos que facilitem sua integração e interoperabilidade, permitindo com que os usuários possam se beneficiar do acesso a essa documentação de maneira simples e objetiva.

Martins, Silva e Carmo (2017) apontam, mapeando as principais iniciativas internacionais de organização de acervos digitais, que se consolida no cenário mundial práticas de gestão de acervos que se integram em rede a partir do compartilhamento e adoção coletiva de padrões e normas que visam publicar informação semântica a partir dos mesmos princípios. Novos modelos de governança têm surgido nesse cenário, produzindo experiências e resultados significativos de valorização da cultura pela adoção de novas práticas de publicação da informação em rede. A agregação dos serviços em pontos unificados de acesso facilita a vida do usuário e valoriza o reuso dos objetos digitais que se tornam acessíveis em uma única interface.

Diversos estudos têm apontado que práticas ligadas a web semântica e aos dados abertos ligados vêm sendo utilizadas como estratégias para interconexão de acervos e aumento da disponibilidade das informações em ambiente de rede para instituições culturais (MARCONDES, 2012; MARCONDES, 2016; CONEGLIAN; SEGUNDO, 2017).

No entanto, os caminhos e as etapas técnicas necessárias para a produção de acervos integrados em rede de forma semântica é ainda algo que carece de pesquisas e da documentação das experiências práticas já realizadas. Há desafios técnicos e organizacionais que precisam ser superados para que a organização dos acervos possa avançar nessa direção. Os problemas vão desde a produção original da documentação dos acervos, na falta de padrões claros estabelecidos e regras de catalogação compartilhadas, até aos softwares utilizados por muitas instituições que ainda não estão preparados para lidar com as informações disponíveis e enriquecidas de forma semântica, além da disponibilização de dados abertos.

Foi visando ampliar a pesquisa a respeito desse cenário que no ano de 2014 o Ministério da Cultura dá início ao projeto Tainacan (MARTINS *et al*., 2017). Agregando o Instituto Brasileiro de Museus (IBRAM) no ano de 2015, o Tainacan vem se tornando um software livre que tem por objetivo facilitar a publicação de acervos digitais em rede das instituições de cultura, sobretudo dos museus. O Tainacan está em processo de implantação em diversas instituições públicas, tais como o Museu Histórico Nacional, o Museu da República, o Museu Villa Lobos, a Fundação Nacional de Artes (FUNARTE), entre outros.

Dessa maneira, no ano de 2017, o Museu do Índio, interessado em participar da rede de instituições que têm experimentado o Tainacan como uma tecnologia de apoio a organização de seus acervos digitais em rede, iniciou um projeto de implantação na instituição. A presente pesquisa visa relatar as etapas metodológicas de execução desse trabalho de implementação, sobretudo, visando compreender quais as etapas técnicas e como elas devem ser encadeadas para potencializar o acesso aos acervos de cultura na direção dos dados ligados e padrões semânticos.

2 DESENVOLVIMENTO

2.1 O acervo digital do Museu do Índio: diagnóstico da situação atual

O museu possui um acervo digital vasto que integra as diferentes áreas da documentação, envolvendo a sua biblioteca, arquivo e museu. No entanto, há diferentes tecnologias e soluções digitais em uso atualmente, fazendo com que tanto a navegação e usabilidade pelos diferentes conteúdos se torne diferente a cada tipologia de acervo, bem como a capacidade de integração dos acervos para a disponibilização de uma busca integrada ao usuário fique comprometida, dificultando o acesso e a compreensão do conjunto da documentação já digitalizada e disponível em rede. Deixa-se de utilizar, dessa maneira, importantes recursos técnicos que tanto podem valorizar como facilitar o acesso ao acervo aos seus usuários de interesse. Apresenta-se no quadro 01 as diferentes tecnologias em uso para os acervos digitais do museu.

Quadro 01: Tecnologias em uso para os acervos digitais do Museu do Índio.

Acervo	Link	Itens	Tecnologia
Museológico	http://base2.museudoindio.gov.br/cgi-bin/wxis.exe?IsisScript=phl82/003.xis&cipar=phl82.cip&acv=002&pft=decorado&lang=por&tmp=/tmp/filecYOOPm	19.865 (documentos)	Software PHL
Bibliográfico	http://base2.museudoindio.gov.br/cgi-bin/wxis.exe?IsisScript=phl82/003.xis&cipar=phl82.cip&acv=003&pft=decorado&lang=por&tmp=/tmp/filecYOOPm	20.208 (documentos)	Software PHL
Arquivístico Bibliográfico	http://www.docvirt.com/docreader.net/docmulti.aspx?bib=museudoindio&pagfis=	623.960 (páginas)	Software Docpro
Arquivístico	http://basearch.museudoindio.gov.br/	0 (documentos)	Software ICA-AtoM

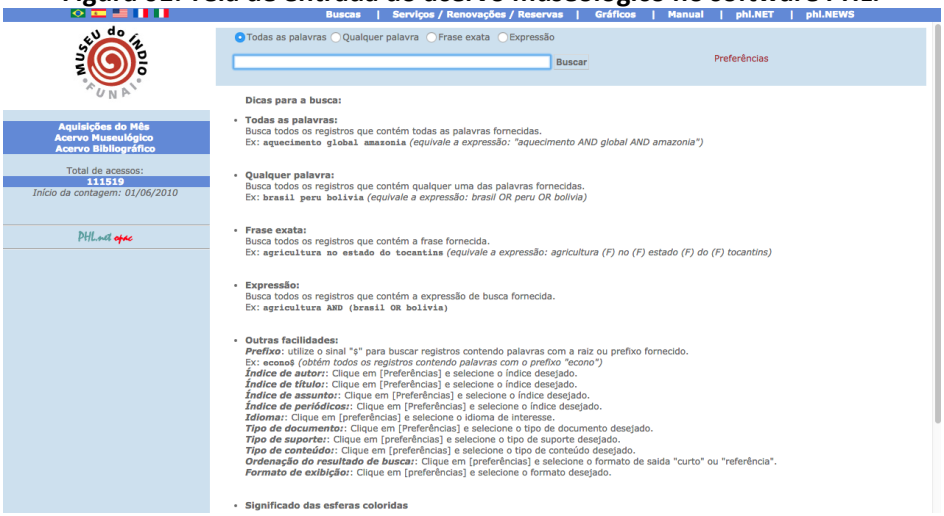
Fonte: Dados da pesquisa.

Vale dizer que para além das tecnologias na quadro 01, o museu utiliza uma outra tecnologia (*software* Joomla[1]) para o gerenciamento de conteúdo do seu site institucional, sendo por meio do acesso a esse ambiente que um usuário consegue localizar os acervos digitais e acessar suas páginas. São, portanto, 4 tecnologias diferentes em uso, gerando 4 bancos de dados diferentes, interfaces gráficas, formas de navegação e modos de organização da informação em questão. A diferença nessas tecnologias inclui a formação dos links de acesso aos conteúdos, sendo que alguns sequer utilizam o domínio institucional do museu (tecnologia Docpro), o que certamente gera impactos na indexação e agregação desse conteúdo em mecanismos de busca na Internet, como Google, Yahoo e Bing, ampliando a dispersão do conteúdo e dificultando a busca integrada mesmo por meio de mecanismos externos aos sistemas de informação do museu.

Outro ponto importante a destacar é o uso de diferentes tecnologias para acervos do mesmo tipo, estando, por exemplo, parte do acervo bibliográfico na tecnologia PHL e Docpro, gerando fragmentações no espaço informacional que são de difícil integração na experiência de uso e no tratamento de dados pelos interessados em pesquisa nesses acervos. O problema inclusive se expressa na forma de contagem da informação, sendo que algumas soluções apontam o número de documentos e outras o número de páginas, dificultando a geração de uma mesma métrica para contabilização dos documentos disponibilizados. Por fim, um outro problema apresentado pelas soluções atualmente em uso é a dificuldade de navegação no acervo por usuários que não tenham uma busca específica a ser realizada e tenham por objetivo explorar o acervo para conhecer as coleções, os conteúdos indexados e compreender melhor as possibilidades de exploração nos dados que as ferramentas ofertam. Para ilustrar, apresenta-se na figura 01 a tela de entrada do acervo museológico no software PHL.

[1] http://www.museudoindio.gov.br/. Acesso em: 23 de jun. 2018.

Figura 01: Tela de entrada do acervo museológico no software PHL.



Fonte: Dados da pesquisa.

Percebe-se na figura 01 que o sistema apenas oferece um campo de busca como alternativa para o usuário encontrar conteúdos no acervo, não havendo a disponibilização de formas de navegação de exploração livre nos documentos digitais, por exemplo, por meio de uma navegação facetada que apresente diferentes perspectivas a respeito dos documentos descritos ou a ordenação dos conteúdos por meio de algum tipo de filtragem, tais como ordem alfabética e data de entrada.

Por fim, os sistemas atualmente em uso no museu não ofertam nenhuma funcionalidade técnica para disponibilidade de dados abertos e ligados a partir das possibilidades de descrição semântica da informação e conexão dos acervos com padrões internacionalmente utilizados. Os dados dos acervos não podem ser baixados localmente nas máquinas dos usuários, dificultando a capacidade de processar esses dados, produzir análises e sínteses da informação que são potencialmente interessantes e necessárias para pesquisadores interessados na identificação de padrões e cruzamentos dos dados do acervo com outras bases de dados.

É considerando esse cenário de alta fragmentação dos acervos, pouca usabilidade na navegação e capacidade de exploração dos conteúdos autonomamente por parte dos usuários, além das poucas possibilidades de consumo dos dados em formatos abertos e semânticos, que o museu se interessa por novas tecnologias que possam suprir essas deficiências ampliando o potencial de uso da informação cultural de alto valor social disponibilizada em seus acervos. É nesse contexto que surge a plataforma Tainacan. Na

presente pesquisa, o trabalho descreve exclusivamente o processo de migração da documentação referente ao acervo museológico.

2.3 Metodologia

A realização de um processo de migração e abertura de uma base de dados envolve diferentes etapas técnicas que precisam ser executadas de forma consequente e articulada. Envolve diferentes procedimentos de tratamento da informação, desde o entendimento das estruturas organizacionais da informação atualmente utilizadas na instituição até os princípios técnicos de conversão dessas estruturas em outras formas de representação e organização da informação, passando pela mudança de padrões técnicos, limpeza, tratamento, normalização, entre outras etapas. Para isso foram realizadas 07 etapas metodológicas de trabalho da informação que são descritas a seguir.

1. **Análise:** realiza-se levantamento das características técnicas dos acervos, procurando identificar padrões de metadados utilizados, políticas de direitos autorais, políticas de digitalização, regras de catalogação em uso, linguagens documentárias, tipologia do acervo, informações técnicas dos softwares, bases de dados e formas de arquivamento dos conteúdos digitalizados. Além disso, esse é um momento fundamental do trabalho de análise e diálogo em conjunto com as instituições acerca do que elas possuem de recursos digitais que podem se tornar dados abertos de interesse público e que possam ser publicados na Internet como coleções visitáveis. É aqui que se identificam catálogos e instrumentos de organização da informação internas da instituição que podem ser potencialmente interessantes ao seu público e que ainda não foram publicados;
2. **Coleta:** os dados de interesse para migração são extraídos da solução atual e disponibilizados em formato tratável pela equipe técnica;
3. **Tratamento:** são realizados procedimentos técnicos de normalização, limpeza, correção de problemas sintáticos e preparação dos dados visando melhoria de desempenho nos processos de busca e recuperação da informação;
4. **Enriquecimento:** são identificados os elementos semânticos na estrutura dos dados que podem ser descritos utilizando ontologias e sistemas de referenciais

amplamente utilizados pela comunidade internacional de dados ligados, tais como DBpedia[2], Wikidata[3], VIAF[4], entre outros;

5. **Migração:** os dados tratados e enriquecidos são migrados para uma nova base de dados no sistema Tainacan, a partir do qual poderá ser publicado na Internet e acessado pelos usuários;
6. **Validação e formação:** os técnicos de documentação do museu navegam, realizam operações de busca e recuperação da informação de maneira a identificar eventuais problemas de migração e identificar necessidades de correção nos dados nas etapas anteriores. Aproveita-se também para formar os técnicos na tecnologia utilizada bem como nos procedimentos técnicos informacionais para manter a nova estrutura e fluxo informacional de gestão das coleções;
7. **Difusão e colaboração em rede:** o acervo é finalmente aberto e disponibilizado para os usuários utilizarem e fornecerem impressões, comentários e colaborarem em diferentes etapas de contribuição aos acervos, a depender dos interesses da instituição em se abrir a esse tipo de diálogo com seus usuários.

Vale dizer que para o presente trabalho serão apresentados apenas os resultados das etapas de 1 a 5 por conta do escopo do artigo e limitação de espaço, ficando as demais para trabalhos futuros.

2.4 Resultados

Os trabalhos junto ao Museu do Índio tiveram início em fevereiro de 2018, quando se realizou a **etapa de análise** dos acervos. Foram realizados dois dias de diálogo com os técnicos do museu, onde importantes recursos informacionais foram identificados. Iniciou-se pela **base de dados** utilizada para registro das informações do acervo museológico, o software PHL. Em relação ao **padrão de metadados**, o museu utilizava para sua catalogação o MARC da maneira como era ofertado pelo software PHL em uso, realizando adaptações de significado semântico dos metadados. Do total de metadados oferecidos, o museu utilizava 19 campos. No quadro 02, pode-se ver os metadados utilizados, o significado semântico atribuído pelo museu aos mesmos, a quantidade de registros identificados com o respectivo

[2] https://wiki.dbpedia.org/. Acesso em 24 jul. 2018.
[3] http://wikidata.org/. Acesso em 24 de jul. 2018.
[4] https://viaf.org/. Acesso em 24 de jul. 2018.

campo catalogado e a taxa relativa de preenchimento do metadado em relação ao total de registros.

Quadro 02: Metadados utilizados e frequência de uso pelos registros da base de dados do Museu do Índio.

Campo	Descrição	Qtd. De Valores	Percentual
v003	Número do Objeto	19717	99,71%
v016	Índio	2860	14,46%
v017	Responsável pela guarda	16072	81,27%
v018	Nome do objeto: Nome técnico da peça	19716	99,70%
v019	Terminologia étnica	4287	21,68%
v025	Coleção/Doador	16580	83,84%
v027	Número de peças	19614	99,19%
v038	Dimensões da peça. (L x A x P)	19448	98,35%
v040	Lingua indígena	17500	88,50%
v043	Nome da etnia	19536	98,79%
v048	Referência Bibliográfica	19553	98,88%
v061	Notas Gerais	2368	11,97%
v066	Estado de origem do objeto	19507	98,64%
v067	País de origem	16848	85,20%
v064	Ano de confecção do objeto	19446	98,34%
v065	Data padronizada	18752	94,83%
v070	Imagem do objeto	8949	45,25%
v071	Categoria	19660	99,42%
v081	Historia administrativa/biografia	1080	5,46%

Fonte: Dados da pesquisa.

Em relação às linguagens documentárias em uso para catalogação, foram identificados dois instrumentos utilizados pelo museu: o **Tesauro de Cultura Material dos Índios no Brasil** (MOTTA e OLIVEIRA, 2006) e o **Dicionário do Artesanato Indígena** (RIBEIRO, 1988). Ambos são utilizados para compor os metadados v048 (referência bibliográfica), v061 (notas gerais) e v071 (categoria). Vale ressaltar que os termos utilizados nos metadados eram inseridos manualmente pelos técnicos do museu a partir da pesquisa nas fontes de informação mencionadas, sem controle terminológico pelo software em uso. Tal recurso de indexação acabou gerando problemas de normalização e padronização dos termos, como se verá a seguir.

Em relação às **regras de catalogação**, a equipe técnica produziu um manual de preenchimento da ficha catalográfica buscando padronizar como cada metadado deveria ser preenchido, facilitando o controle sintático e semântico do trabalho de catalogação. Esse material se mostrou de fundamental importância para o processo de coleta da base de dados, considerando que os técnicos puderam checar as informações coletadas e revisar se

as mesmas apresentavam as informações esperadas conforme o manual indicava, facilitando o reparo de erros técnicos de importação e exportação de dados.

Em relação à **política de digitalização**, o museu possui um amplo trabalho de tratamento do acervo. Dos 19.775 itens do acervo identificados na base de dados, 8.949 (45,3%) já se encontrava com referência direta aos arquivos digitais em alta definição dos objetos. O museu estima que 50% do seu acervo se encontra digitalizado.

A respeito da tecnologia de base de dados e arquivamento da informação, identificou-se que o sistema PHL não apresentava base de dados relacional possível de coletar na infraestrutura de servidores do museu, dado que seu servidor de banco de dados é o WWWisis©Bireme. Tal fato apresentou desafio técnico para a **etapa de coleta** da informação, considerando que os dados tiveram de ser exportados em XML a partir de uma funcionalidade técnica do software. A exportação resultou em dois arquivos em formato texto estruturados no padrão XML com as informações dos 19 campos utilizados pelo museu. Os arquivos possuíam o tamanho de 51,4MB e 4,7MB, contendo o primeiro os metadados descritivos e o segundo os administrativos.

Já na **etapa de tratamento**, os dados coletados foram analisados para verificar sua estrutura e a adequação dos arquivos XML para posterior conversão. Utilizou-se o software XML Viewer Plus[5] para analisar as eventuais inconsistências estruturais dos arquivos. Foi utilizada a função **validar**, que verifica se todas as *tags* XML estão corretamente dispostas e se existe algum caractere que cause interferência na estrutura do documento, tais como caracteres especiais que podem provocar quebra de linha e prejudicar uma etapa de conversão de dados. Os problemas encontrados tinham três características em comum e suas respectivas soluções: 1 – *tags* de links mal formadas "<Link>Descrição" foram organizados, onde faltou a abertura ou fechamento da *tag* foi corrigido; 2 – caractere "&" com interferência no documento foram substituídos por “e”; 3 – caracteres "<", ">" no meio de textos sem indicação de *tags*, com interferência na leitura do XML foram removidos. Entende-se que os problemas encontrados se devem a dificuldade do sistema antigo em exportar seus dados, já que alguns dos caracteres inseridos nas informações de caracterização dos objetos do museu foram confundidos com *tags* ou interferiram na leitura

[5] XML Viewer Plus – http://www.alexnolan.net/software/freexmleditor.htm . Acesso em 24 de jul. 2018.

do documento, além disso, os links das imagens foram exportados com falhas de abertura ou fechamento de *tags*, impedindo que os dados fossem obtidos corretamente.

A partir da validação dos arquivos XML foi desenvolvido um código de programação na linguagem Python[6], para ler especificamente a estrutura do arquivo XML exportado pelo PHL e estruturar os dados no formato tabular, onde as *tags* de atributos são transformadas em colunas e os valores dentro dessas *tags* armazenados em células.

Uma vez os arquivos corrigidos estruturalmente e convertidos em formato de fácil tratamento, adotou-se o software OpenRefine[7] para analisar eventuais **problemas de normalização, reconciliação e desambiguação** dos termos de indexação e explorar os demais metadados em busca de se identificar padrões que poderiam ser facilmente corrigidos de maneira semiautomática pelas funcionalidades disponíveis no *software*. Foram utilizadas várias estratégias para tratamento dos dados, visando, sobretudo, melhorar a qualidade da documentação já existente em relação ao acervo do museu e facilitar os processos de busca e recuperação da informação. As técnicas utilizadas para cada campo de metadados descritos no quadro 02 foram: limpeza de espaços em branco iniciais, remoção de espaços em branco consecutivos, padronização de separador de textos (havia diferentes formatos, tais como ",", ";","/" entre outros), normalização dos termos em caixa baixa, aplicação dos algoritmos de agrupamento que, segundo Stephens (2018), "encontram grupos de diferentes valores que são formas alternativas de representação da mesma coisa", transferência de metadados para os campos corretos quando foram encontradas ocorrências em campos trocados, separação de metadados para compor campos múltiplos quando foi identificado que duas ou mais informações estavam agrupadas de forma incorreta em um campo e, por fim, reconciliação para dados normalizados (caso de nomenclatura das etnias indígenas, onde houve normalização das entradas). No quadro 03, observa-se o resultado do trabalho de tratamento de dados produzido com o OpenRefine, ressaltando a porcentagem de redução de termos obtida após a execução de todas as técnicas listadas acima.

[6] Link do Código disponível em: https://github.com/LuisMDlab/projeto_base_museus/edit/master/xml_to_Csv.py. Acesso em 24 jul. 2018.
[7] Open Refine – www.openrefine.org. Acesso em 24 jul. 2018.

Quadro 03: Redução do número de termos dos metadados do acervo museológico do Museu do Índio.

Código MARC	Metadado	Nº de Termos base de dados inicial	Nº de Termos base de dados normalizada	% de redução de termos
v094	Estado de conservação	2236	3	-99,87%
v017	Responsável pela Guarda	61	2	-96,72%
v048	Referência Bibliográfica	662	54	-91,84%
v098	Instituição detentora	10	1	-90,00%
v066	Estado de Origem do Objeto	132	24	-81,82%
v087	Descritor temático	5877	1733	-70,51%
v095	Materia Prima	5744	1937	-66,28%
v093	Técnica de confecção	757	301	-60,24%
v067	Páis de Origem	15	6	-60,00%
v064	Ano de Confecção do Objeto	478	226	-52,72%
v099	Intervenções	2	1	-50,00%
v043	Nome da Etnia	359	265	-26,18%
v081	História Administrativa/biografia	249	194	-22,09%
v092	Quantidade de Partes	47	39	-17,02%
v016	Nome do Artesão	942	798	-15,29%
v018	Nome do objeto	1574	1351	-14,17%
v089	Observações1	1981	1750	-11,66%
V082	Localidade	883	782	-11,44%
v083	Descrição do objeto	15066	14367	-4,64%
V086	Descreve-se brevemente a função da peça descrita pela etnia	3770	3677	-2,47%
V061	Notas Gerais	963	947	-1,66%
v038	Dimensões da Peça	14321	14238	-0,58%
v065	Data Padronizada	774	772	-0,26%
v03	Número de registro do objeto	1958	1956	-0,10%
v040	Língua Indígena	45	45	0,00%
v019	Terminologia Étnica	1437	1716	19,42%
v088	Descritores secundários	207	285	37,68%

Fonte: Dados da pesquisa.

É importante ressaltar que há 11 metadados nos quais se obteve uma redução de 50% ou mais no número de termos utilizadas na catalogação dos objetos. É um resultado de destaque, uma vez que essa redução representa a resolução de problemas técnicos sintáticos da informação descritiva e agrupa conceitos que representavam a mesma informação mas estavam escritos de formas ligeiramente diferentes, a ponto de serem identificados de forma semiautomática pelos algoritmos de agrupamento. Na média, pode-se dizer que houve uma redução de 17% dos termos da base de dados de documentação como um todo. Esse resultado traz impactos na capacidade de melhoria de uso do sistema pelos usuários, que passam a ter maior precisão e confiabilidade nos resultados quando executam ações de busca e recuperação da informação. Vale ressaltar que para as duas últimas ocorrências do quadro 03, a terminologia étnica e descritores secundários, houve

aumento do número de termos por conta da separação de termos que estavam anteriormente agrupados, permitindo com que eles fossem descritos em campos múltiplos e não todos juntos em um mesmo campo de texto. Entende-se que essa etapa de tratamento representa uma contribuição técnica de alta relevância para a discussão dos procedimentos técnicos da documentação museológica para área.

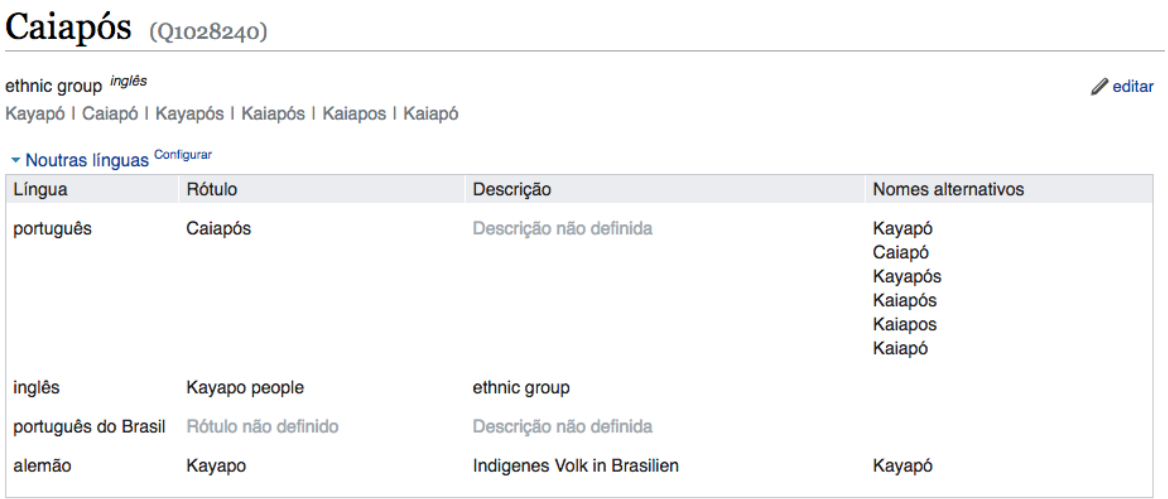
Uma vez a informação tratada tecnicamente na etapa descrita acima, passou-se a **etapa de enriquecimento**. A importância dessa etapa consiste no trabalho de identificar as possíveis formas de ligar os dados em trabalho com outras fontes de informação, ampliando o valor de uso da informação na medida em que conecta os registros em redes semânticas a partir de padrões públicos e amplamente utilizados. Novamente, utilizamos o software OpenRefine a partir da funcionalidade de reconciliação com a base semântica Wikidata para essa etapa. Entende-se que a Wikidata vem se transformando numa importante referência internacional para interconectar diversos sistemas de indexação e controle de autoridades semânticos atualmente em uso (ERXLEBEN *et al*., 2014). Há importantes iniciativas no mundo dos acervos museológicos se utilizando da mesma estratégia, tendo inclusive por protagonista a iniciativa de agregação de acervos digitais de museus, galerias, bibliotecas e arquivos da União Européia, a Europeana, que realizou no ano de 2017 uma chamada pública (EUROPEANA, 2017) incentivando as instituições a publicarem seus vocabulários na Wikidata, entendendo que ações como essa poderiam trazer benefícios a comunidade de instituições guardiãs de acervos no compartilhamento e reconciliação de seus vocabulários, beneficiariam a Wikidata, pois possibilitaria o enriquecimento dessa base de conhecimento público tornando-a mais robusta e por fim, beneficiariam a Europeana por possibilitar o acesso a conhecimento mais especializado sobre sobre os seus acervos por meio da wikidata. Dessa forma, entende-se a Wikidata como um dos maiores agregadores disponível publicamente para reconciliação com vocabulários controlados dos mais diferentes tipos, envolvendo desde conceitos de áreas específicas de conhecimento, como o caso de uma pesquisa sobre genética que se vale da plataforma como uma importante fonte de desambiguação de conceitos (BURGSTALLER-MUECHLBACHER, 2016), e até o controle de autoridades de nomes de organizações, países, artistas, entre outros.

A aplicação da etapa de enriquecimento na base do Museu do Índio se deu, como etapa inicial de experimentação, a partir do metadado v043-Nome da Etnia. O objetivo era encontrar na Wikidata os códigos que representassem o nome das etnias indígenas e

dessem acesso aos demais conteúdos ali disponíveis, tais como outras propriedades a respeito das etnias: variações na forma de escrita do nome, tamanho da população, geolocalização, endereço das páginas na Wikipédia nas diferentes línguas que descrevem verbetes para a respectiva etnia, entre outros. Vale dizer que a Wikidata é um sistema que lida com a informação de maneira multilíngue, facilitando com que se possa encontrar a mesma informação descrita e representada em várias linguagens a partir de um único código de controle do conceito representado. Dessa maneira, esperava-se, por meio desta etapa, ter acesso a rede de descrição de verbetes das etnias indígenas em várias línguas, enriquecendo a documentação do museu com essa diversidade de informações e referências para uma rede especializada de informação técnica, além de identificar um metadado da documentação do museu com um identificador reconhecido internacionalmente. Esse é um passo fundamental na direção da abertura da documentação no formato dados ligados semânticos, etapa a ser realizada em pesquisa futura, mas que já se preparava nesse momento de enriquecimento da informação.

Como exemplo, observa-se nas figuras 02 (nomes alternativos), 03 (outros identificadores internacionais que se referem a esse mesmo conceito) e 04 (verbetes de diferentes línguas disponíveis na Wikipedia sobre o conceito) informações disponíveis na Wikidata sobre a etnia Caiapó, identificador Q1028240, que, uma vez identificada na base do museu, dá acesso ao enriquecimento de todas essas informações na base.

Figura 02: Nomes alternativos ao conceito Caiapós.



Caiapós (Q1028240)

ethnic group *inglês* editar

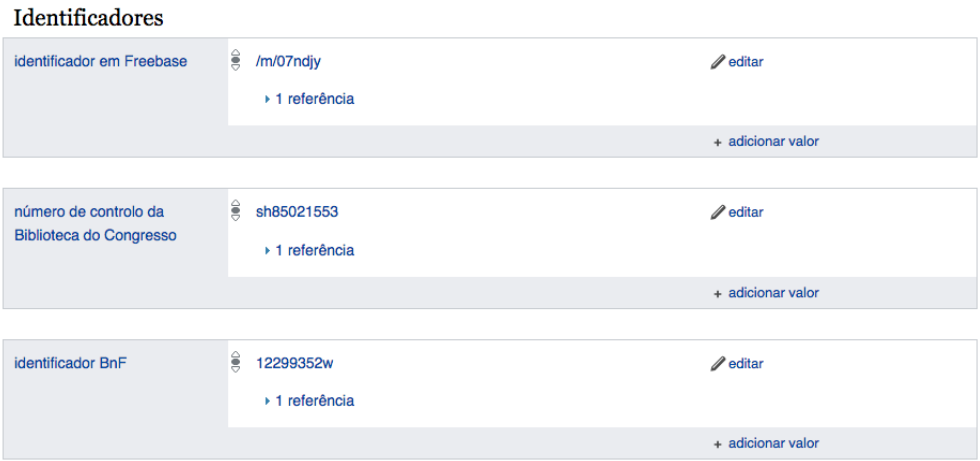
Kayapó | Caiapó | Kayapós | Kaiapós | Kaiapos | Kaiapó

Noutras línguas Configurar

Língua	Rótulo	Descrição	Nomes alternativos
português	Caiapós	Descrição não definida	Kayapó Caiapó Kayapós Kaiapós Kaiapos Kaiapó
inglês	Kayapo people	ethnic group	
português do Brasil	Rótulo não definido	Descrição não definida	
alemão	Kayapo	Indigenes Volk in Brasilien	Kayapó

Fonte: Wikidata.

Figura 03: Outros identificadores internacionais representando o conceito Caipós.



Fonte: Wikidata.

Figura 04 – Verbetes em várias línguas descrevendo o conceito Caiapós na Wikipedia.

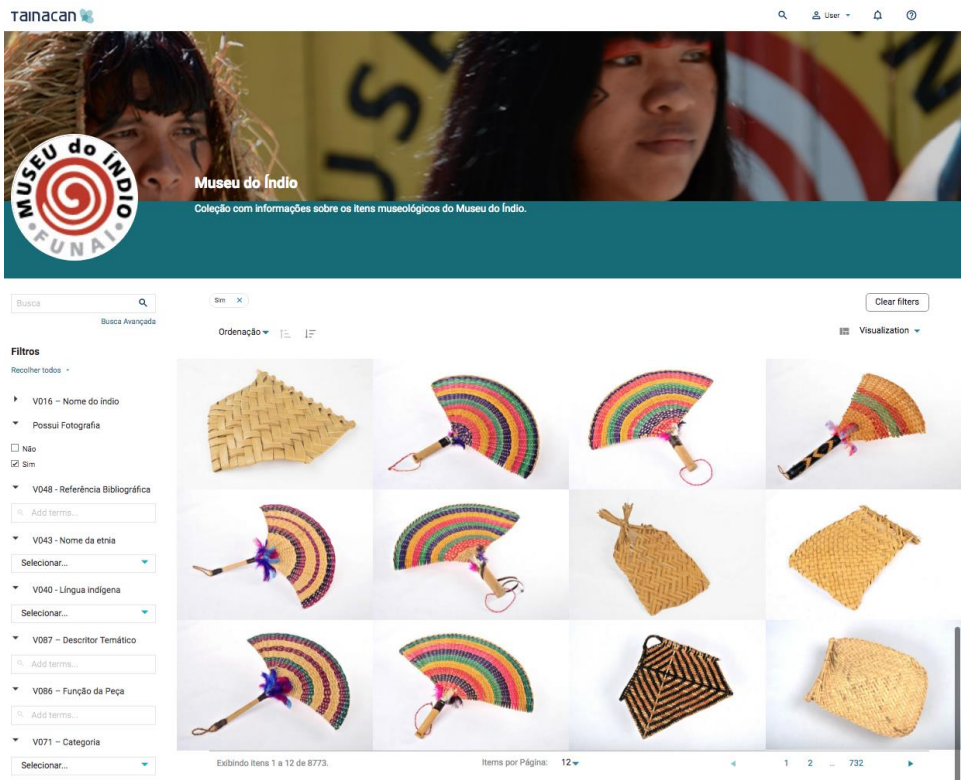
Fonte: Wikidata.

Ao produzir a conexão da base do museu com a Wikidata, passa-se a ter acesso a toda essa informação disponível, que pode ser incorporada em metadado específico na documentação do acervo e apresentada ao usuário quando de seu acesso aos registros. Vale ressaltar que na figura 03, pode-se também utilizar esses identificadores internacionais para fornecer outros links a outras bases de dados que possam conter informações relevantes ao conceito representado. Do total de 265 termos de nomes de etnias existentes na base, conforme dados do quadro 03, encontrou-se na Wikidata 232 (87,5%), permitindo que uma taxa bastante elevada de etnias pudessem ser enriquecidas com as referências apresentadas

pela Wikidata, conforme se apresentou nas figuras acima. Entende-se que essa contribuição à documentação do museu o coloca em outro paradigma informacional, fazendo com que se torne de fato uma informação interoperável em rede, permitindo dialogar com bases de controle de autoridades internacionais, abrindo as portas para que novos produtos e serviços informacionais possam ser gerados a partir dessa conexão.

Finalizados os trabalhos de enriquecimento dos dados, passou-se à **etapa de migração**, última etapa a ser apresentada na presente pesquisa, quando os mesmos são transferidos para o *software* livre Tainacan e disponibilizados em ambiente web. Os dados foram importados utilizando-se a API pública do Tainacan, a partir da configuração prévia dos tipos de metadados a serem utilizados em uma nova coleção. Além disso, foram configuradas as facetas que serviriam de estratégia de navegação e recuperação da informação do acervo. As informações técnicas sobre como proceder a instalação e migração dos dados realizadas nesta etapa podem ser encontradas em Martins *et al.* (2017). Apresenta-se na figura 05 o resultado final da migração e disponibilização da documentação do acervo no *software* livre Tainacan.

Figura 05: Documentação do Museu publicada no *software* livre Tainacan.



Fonte: Dados da pesquisa.

Observa-se na figura 05 algumas características da interface gráfica proposta pelo Tainacan: as facetas para filtro e navegação ao lado esquerdo da tela, o acesso aos objetos digitais no corpo central da página, diferentes funcionalidades de visualização dispostas na barra de cima dos objetos, tais como a forma de ordenação dos itens, a forma de visualização e a barra de busca simples e avançada. Com isso, conclui-se a descrição dos resultados obtidos.

3 CONSIDERAÇÕES FINAIS

O presente trabalho apresentou os resultados obtidos na realização de 05 etapas técnicas de preparação da documentação do acervo museológico do Museu do Índio para sua disponibilização na Internet por meio do software livre Tainacan. O trabalho teve como preocupação demonstrar em detalhes os procedimentos técnicos realizados bem como os resultados obtidos em cada etapa visando, sobretudo, ressaltar a importância da realização desse tipo de trabalho para que se possa obter resultados de melhor qualidade da informação quando de sua disponibilização em ambiente de rede para acesso público.

Expressivos resultados foram obtidos na etapa de tratamento de dados, quando se pode reduzir os problemas sintáticos, má formação e erros de catalogação presentes na documentação. Tal etapa terá efeito significativo na realização de tarefas de busca e recuperação da informação, melhorando precisão e a qualidade dos resultados, permitindo aos usuários encontrar corpus de documentos mais precisos. Espera-se investigar tais efeitos em trabalhos futuros.

Já na etapa de enriquecimento, o resultado obtido de mais de 85% de conexão dos nomes das etnias indígenas com a plataforma Wikidata traz não apenas resultados imediatos de enriquecimento da documentação museológica, mas gera a possibilidade da produção de novos serviços e produtos informacionais. De imediato, a informação apresentada no Tainacan pode oferecer os links encontrados para as páginas na Wikidata e na Wikipedia referentes a cada etnia indígena nas várias línguas presentes, permitindo com que os usuários possam navegar nessa informação de maneira sistematizada a partir da documentação do museu. É um complemento que facilita e complementa os recursos de pesquisa ofertados pelo acervo digital. Como perspectiva futura, tal ação permite ao museu se conectar a Wikidata de várias formas, por exemplo, enviando parte do seu acervo para ser disponibilizado na Wikimedia Commons, tornando-se disponível para ilustrar verbetes

das etnias indígenas das várias línguas presentes na Wikipedia, ampliando a capacidade de difusão e reuso dos conteúdos digitais disponíveis em seu acervo. Além disso, essa estratégia aponta um caminho para integração dos próprios acervos internos da instituição, agregando biblioteca, arquivo e museu a partir do compartilhamento das mesmas referências semânticas controladas por meio da Wikidata. Pretende-se explorar em pesquisas futuras outras possibilidades de conexão do acervo digital do museu disponível no Tainacan com as soluções Wikimedia.

Por fim, o Tainacan se mostra uma ferramenta que apresenta grande potencial de oferecer solução aos principais problemas apresentados na seção de diagnóstico da situação atual do museu. Em relação ao cenário de alta fragmentação dos acervos o software permite integrar diferentes coleções e formatos de documentação, em relação a pouca usabilidade na navegação do software atual e a baixa capacidade de exploração dos conteúdos autonomamente por parte dos usuários, o Tainacan oferta interface com navegação facetada e diferentes formas de visualização da informação, por último, em relação às poucas possibilidades de consumo dos dados em formatos abertos e semânticos, a solução atual apresentou um caminho simples, mas que se vale da capacidade de reconciliar dados da documentação do museu com a Wikidata e apresentá-los ao usuário no Tainacan como informação incorporada digitalmente em metadado específico e modelado para esse fim.

Entende-se que esse caminho técnico de trabalho com a documentação relativa ao acervo do museu, visando sua migração e abertura para acesso público na web, amplia seu valor de uso e sua capacidade de se socializar em rede, apoiando o trabalho social e político que compete ao Museu do Índio na preservação e promoção do patrimônio cultural indígena.

REFERÊNCIAS

BURGSTALLER-MUECHLBACHER, Sebastian *et al.* Wikidata as a semantic framework for the Gene Wiki initiative. **Database:** The Journal of Biological Databases and Curation, v.2016, Mar. 2016. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795929/pdf/baw015.pdf> Acesso em: 24 jul. 2018.

CONEGLIAN, C. S.; SEGUNDO, J. E. S. Europeana no linked open data: conceitos de web semântica na dimensão aplicada das humanidades digitais. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, v.22, n.48, 2017. Disponível em: <http://www.brapci.inf.br/v/a/22274>. Acesso em: 06 Mai. 2018.

ERXLEBEN, Fredo *et al.* Introducing Wikidata to the linked data web. In: INTERNATIONAL SEMANTIC WEB CONFERENCE. **Proceedings…** Springer, Cham, 2014. p.50-65. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-11964-9_4>. Acesso em: 24 jul. 2018.

EUROPEANA. **Get your vocabularies in Wikidata... so Europeana and others can get them**. Disponível em: <https://pro.europeana.eu/page/get-your-vocabularies-in-wikidata>. Acesso em 21 jul. 2018.

MARCONDES, Carlos Henrique. Linked data - dados interligados - e interoperabilidade entre arquivos, bibliotecas e museus na web. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, v.17, n.34, p.171-192, 2012. Disponível em: <http://www.brapci.inf.br/v/a/12657>. Acesso em: 06 maio 2018.

_____. Interoperabilidade entre acervos digitais de arquivos, bibliotecas e museus: potencialidades das tecnologias de dados abertos interligados. **Perspectivas da Ciência Informação,** Belo Horizonte, v.21, n.2, p 61-83, jun. 2016. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362016000200061&lng=en&nrm=iso>. Acesso em: 06 maio 2018.

MARTINS, Dalton Lopes; SILVA, Marcel Ferrante; CARMO, Danielle do. Acervos em rede: perspectivas para as instituições culturais em tempos de cultura digital. **Em Questão**, v.24, n. 1, 2018. Disponível em: <http://www.brapci.inf.br/v/a/29898>. Acesso em: 22 jul. 2018.

MARTINS, Dalton Lopes *et al.* Repositório digital com o software livre Tainacan: revisão da ferramenta e exemplo de implantação na área cultural com a revista Filme Cultura. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais...** Marília: UNESP, 2017. Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/472/838 >. Acesso em: 21 jul. de 2018.

MOTTA, Dilza Fonseca da, OLIVEIRA, Leandra de. **Tesauro de Cultura Material dos Índios no Brasil**. [S.l.]: Museu do Índio; FUNAI, 2006. 249p.

RIBEIRO, Berta G. **Dicionário do artesanato indígena**. Belo Horizonte: Itatiaia; São Paulo : EDUSP, 1988. 352p.

MUSEU DO INDIO. **O Museu**. Disponível em: <http://www.museudoindio.gov.br/o-museu/apresentacao> Acesso em: 21 julho 2018.

STEPHENS, Owen. **Clustering in Depth:** methods and theory behind the clustering funcionality in OpenRefine. Disponível em: <https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth>. Acesso em: 21 jul. 2018.