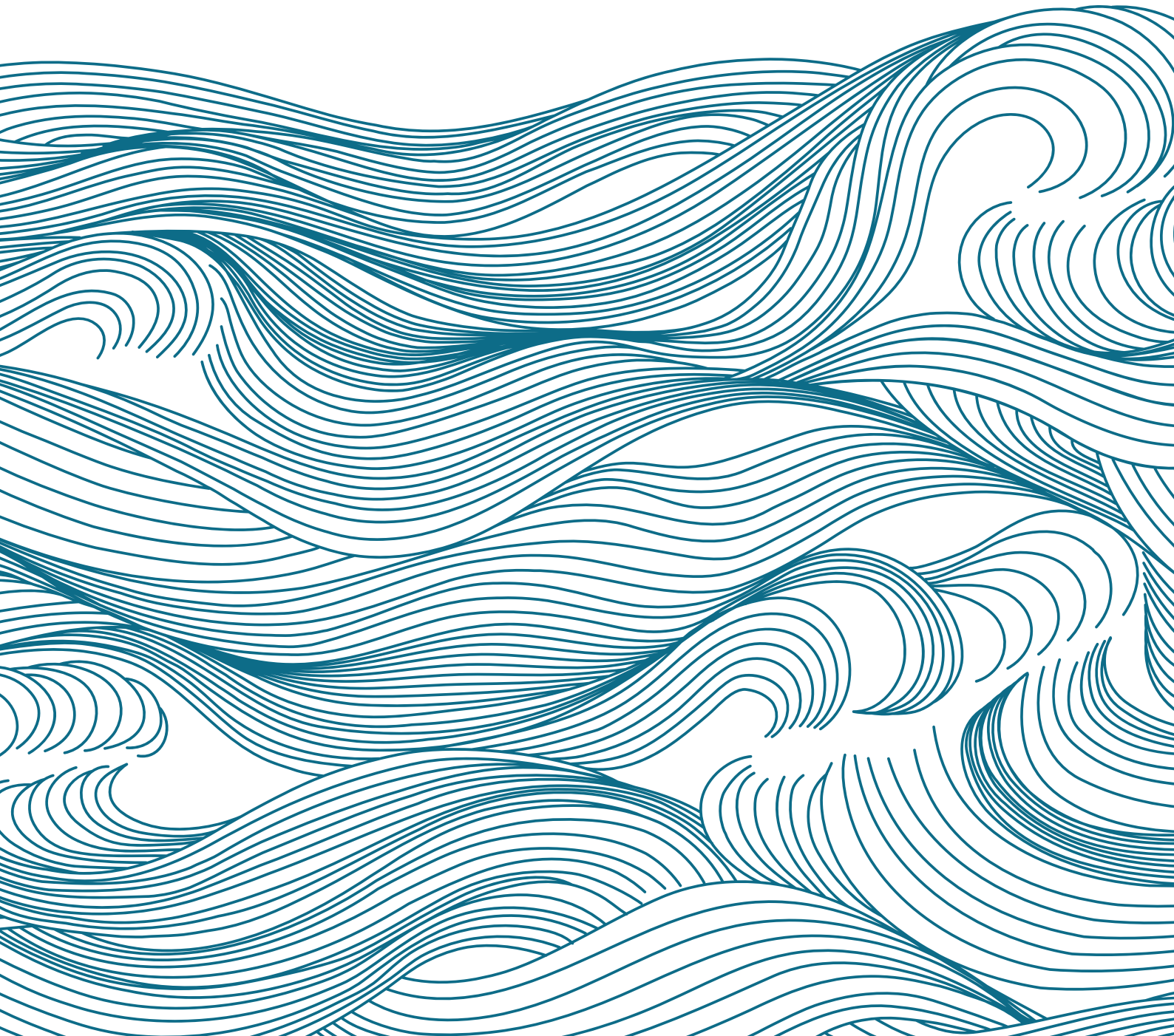
MEMÓRIA EM REDE

RELATÓRIO DE CUMPRIMENTO META 03

ESTUDOS SOBRE ALIMENTAÇÃO AUTOMÁTICA DE PASSIVO



RELATÓRIO DE CUMPRIMENTO META 03

ESTUDOS SOBRE ALIMENTAÇÃO AUTOMÁTICA DE PASSIVO

ibict Instituto Brasileiro de Informação em Ciência e Tecnologia
UNIDADE DE PESQUISA DO MCTI

Brasília

2024

PRESIDENTE DA REPÚBLICA
Luiz Inácio Lula da Silva

VICE-PRESIDENTE DA REPÚBLICA
Geraldo José Rodrigues Alckmin Filho

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
Luciana Santos
Ministra da Ciência, Tecnologia e Inovação

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA

Tiago Emmanuel Nunes Braga
Diretoria

Carlos André Amaral de Freitas
Coordenação de Administração - COADM

Ricardo Medeiros Pimenta
Coordenação de Ensino e Pesquisa em Informação para a Ciência e Tecnologia - COEPI

Henrique Denes Hilgenberg Fernandes
Coordenação de Planejamento, Acompanhamento e Avaliação - COPAV

Cecília Leite Oliveira
Coordenação-Geral de Informação Tecnológica e Informação para a Sociedade - CGIT

Washington Luís Ribeiro de Carvalho Segundo
Coordenação-Geral de Informação Científica e Técnica - CGIC

Hugo Valadares Siqueira
Coordenação-Geral de Tecnologias de Informação e Informática - CGTI

Milton Shintaku
Coordenação de Tecnologias para Informação - COTEC

Ministério da Ciência, Tecnologia e Inovações
Instituto Brasileiro de Informação em Ciência e Tecnologia

RELATÓRIO DE CUMPRIMENTO META 03

ESTUDOS SOBRE ALIMENTAÇÃO AUTOMÁTICA DE PASSIVO

Brasília
2024

EQUIPE TÉCNICA

Diretor do Instituto Brasileiro de Informação em Ciência e Tecnologia
Tiago Emmanuel Nunes Braga

Coordenador-Geral de Tecnologias de Informação e Informática – CGTI
Hugo Valadares Siqueira

Coordenador do Projeto
Milton Shintaku

Autores
Danielle do Carmo
Diego José Macedo
Graziela Barros Gomes
Gustavo Cardoso Paiva
João Maricato
Maison Roberto Mendonça Gonçalves
Milton Shintaku
Paulo Henrique
Thayane Alencar

Revisão
Rafael Teixeira de Souza

Normalização
Maison Roberto Mendonça Gonçalves

Diagramação e projeto gráfico
Rafael Fernandez Gomes

R382 Relatório de cumprimento da meta 03: estudos sobre alimentação automática de passivo / Milton Shintaku ... [et al.] Brasília: Ibict, 2023.
28 p.

1.assunto. 2.Bens culturais 3. Inventário Nacional de Referências Culturais. I. Shintaku, Mil ton. II. Instituto Brasileiro de Informação em Ciência da Informação. III. Instituto do Patrimônio Histórico e Artístico Nacional.

CDU 002:004
CDD 303.4833

Ficha catalográfica elaborada por Maison Roberto Mendonça Gonçalves CRB10/2689

Este Relatório de Técnico é um produto do Projeto de Pesquisa intitulado Organização e Difusão dos Acervos Digitais do Patrimônio Cultural: A Memória em Rede em parceria do Instituto do Patrimônio Histórico Artístico Nacional (IPHAN).

Ref. IBICT - Processo SEI nº 01302.000297/2022-95
Ref. IPHAN - Processo SEI nº 01450.003123/2020-19

Sumário

1. INTRODUÇÃO

O termo "repositório" tem assumido acepções amplas no contexto dos sistemas de informação. Originalmente, com o movimento de acesso aberto (*Open Access*), era entendido como sistema voltado para a obtenção de cópias de artigos científicos publicados em revistas pagas, como uma segunda fonte (Weitzel, 2006). Por isso, foi denominada por Harnad *et al.* (2004) como via verde (green road), ou seja, um sistema que dava "sinal verde" ao acesso ao conhecimento científico, o qual antes restringia-se aos assinantes.

Posteriormente, os repositórios passaram a ser conhecidos como um sistema que facilita o acesso à documentação, inclusive publicando primeiras fontes (Shintaku; Vidotti, 2016). Assim, eles passaram a fomentar o acesso à literatura cinzenta como as teses e dissertações, que não passaram por processo editorial tradicional. Com isso, em algumas áreas, repositórios passaram a ser considerados sistemas de informação que dão acesso a um acervo documental em variada tipologia e formatos.

Com base nesse cenário, podem ser apontadas diversas tecnologias para atendimento aos preceitos dos repositórios. Possivelmente, o mais conhecido e utilizado sistema para criação de repositórios, conforme as orientações acadêmicas, é o Dspace, criado no *Massachusetts Institute of Technology* (MIT). De forma mais ampla, outras tecnologias podem ser utilizadas para serem repositórios, tais como o Omeka, Tainacan, CKAN e Archivematica, cada qual com as suas especificidades, voltadas para o atendimento das necessidades dos usuários, indo além de acervos e documentos acadêmicos.

Nesse contexto, no projeto de pesquisa firmado entre o Instituto do Patrimônio Histórico e Artístico Nacional (Iphan) e o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), denominado de *Organização e difusão dos acervos digitais do patrimônio cultural: a memória em rede*, teve por objeto a alimentação do sistema do Novo Inventário Nacional de Referências Culturais (INRC) e do Banco de Bens Culturais Registrado (BCR). Como sistema de repositório para os dois conjuntos de acervos foi escolhido o Tainacan, *software* livre, baseado no *Content Management System* (CMS) Wordpress.

O estudo relatado no presente relatório é caracterizado como pesquisa aplicada, com significativo uso de metodologias ligadas à Ciência da Computação e à Ciência da Informação. Trata-se de um estudo experimental, empírico, com viés teórico e prático, que utilizou métodos quantitativos e qualitativos. Há destaque para técnicas de mineração de texto, tais como:

- Extração de textos e imagens utilizando bibliotecas de leitura de arquivos em pdf como PyMuPDF/fitz e PDFminer.

- Processamento de linguagem natural com bibliotecas como Spacy e NLTK para remover palavras indesejadas que não ajudavam na análise e categorização e encontrar padrões específicos como nome de pessoas em um documento.

- ORC, utilizando o *software* Tesseract para ler, extrair e inserir textos em documentos digitalizados.

- Desenvolvimento e utilização de expressões regulares para identificar padrões de texto e coletar itens específicos.

No ambiente da pesquisa desenvolvida, observou-se que as fontes de informação são um conjunto de documentos digitalizados e nativos digitais em variados formatos (texto, imagem, som, etc), grande volume de dados com baixo nível de padronização documental e de definição de metadados. Assim, tornou-se necessário o desenvolvimento de diversos processos visando a sua importação automatizada para o *software* Tainacan. Os programas estão disponíveis no GitHub1, onde são compartilhados os códigos-fonte, viabilizando a sua reutilização.

1 https://git.ibict.br/publico/mineracao-de-textos-do-irnc-e-rbcnis

2. OBJETIVOS

O estudo tem como objetivo geral a realização de uma investigação para alimentação automática da documentação passiva, ou seja, dos acervos legados do INRC e do BCR.

Com essa intenção, foi realizada uma série de etapas para o alcance dos objetivos específicos a seguir, aplicando-os aos dois tipos de acervo em relação aos seus sistemas de repositório em questão: Estudos com algoritmos de ciência de dados para tratamento de dados; e Elaboração de procedimentos e carga automatizada de documentos.

3. RESULTADOS

3.1 Estudos com algoritmos para tratamento de dados

Na primeira etapa foi realizada a análise documental e identificação de metadados por meio da análise qualitativa visando a identificação de padrões passíveis de extração automatizada. Os metadados identificados foram incluídos no Tainacan para posterior importação dos dados.

A segunda etapa consistiu na categorização dos documentos por meio da análise dos conteúdos dos documentos. Tal etapa foi necessária em razão da variedade terminológica dos nomes dos arquivos. A categorização proposta foi ajustada e validada com a utilização dos métodos computacionais e manualmente pelos especialistas do Iphan e do Ibict.

Na terceira etapa realizou-se a padronização do título dos arquivos na sua renomeação automatizada (realizada a partir das categorias desenvolvidas na etapa 2).

A quarta etapa consistiu na conversão dos arquivos em imagem e em texto, sendo realizada a conversão dos documentos para imagens, a fim de possibilitar a posterior extração dos dados em forma de texto. Essa etapa foi necessária, pois existiam arquivos digitalizados em diversos formatos, inviabilizando a extração de dados textuais por meio de reconhecimento óptico de caracteres (OCR).

Na quinta etapa foi realizada a mineração dos textos, os quais foram extraídos dos documentos utilizando o processo de OCR e de um *script* criado na linguagem python. Os dados foram, então, salvos em um arquivo no formato csv no modelo aceito para importação pelo Tainacan.

A sexta e última etapa contemplou a importação e visualização dos dados no *software* Tainacan. Tal etapa foi realizada por meio da ferramenta de importação, disponível no *software*, a qual possibilita a criação de itens em massa de diferentes coleções e a criação automática de uma página do item que exibe seus dados, documentos e mídias associadas.

3.2 O acervo Legado do BCR

A base documental do BCR foi disponibilizada pelo Iphan em um ambiente do Google Drive. O acervo do BCR corresponde a 1.027 arquivos, sendo 637 documentos de texto e 390 mídias (fotografias e vídeos). Foram realizados estudos e procedimentos para a inserção no Tainacan desses 1.027 documentos.

Os diferentes tipos de documentos do BCR foram analisados manualmente, buscando-se identificar dados passíveis de mineração. Foram realizados testes visando à extração de dados em alguns deles. No entanto, o único

documento que continha dados com determinados padrões (possibilitando a mineração) foi o denominado "Certidão de registro de bem Cultural".

O Iphan definiu previamente que os metadados para os RBCNIs no Tainacam fossem: Título do bem cultural; Descrição; Abrangência do registro; Território já identificado; Localização; Livro de Registro; Instituições parceiras; e, Documentos; Mídias; Data de Registro; Link para o processo SEI - Registro; Data de Revalidação; e Link para processo SEI -Revalidação.

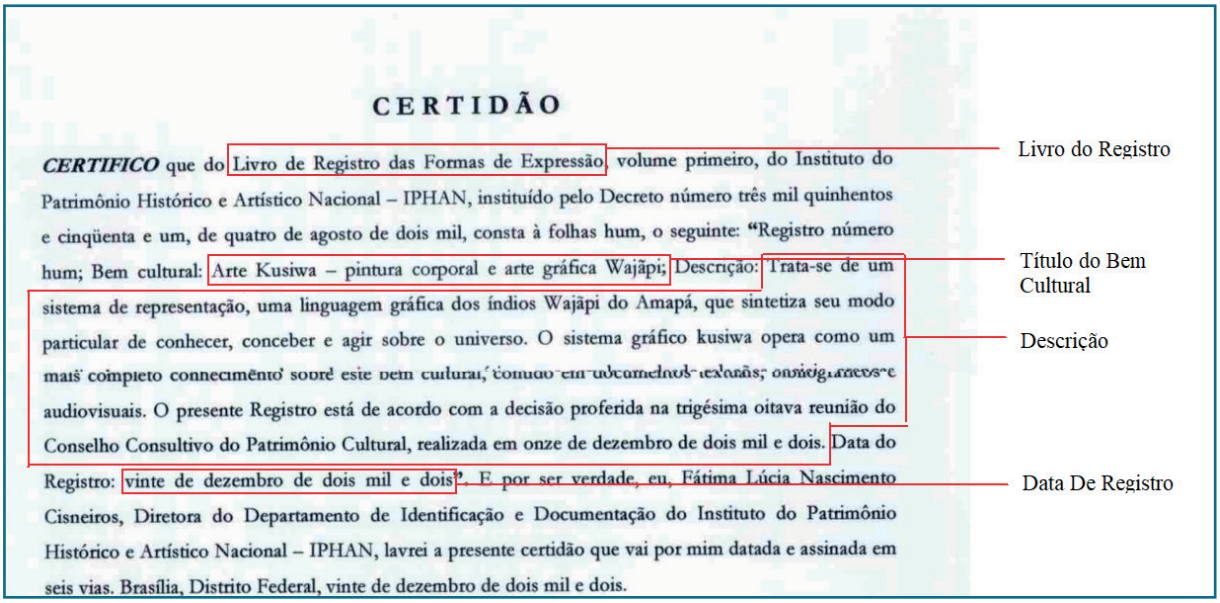


Figura 1 - Exemplo ilustrativo de documento analisado e a definição de dados minerados e inseridos nos metadados no Tainacan

Fonte: Elaborado a partir de documento do Iphan (2023)

Assim, os dados **"Livro de registro"**, **"Título do bem cultural"** e **"Descrição"** e **"Data de registro"** foram definidos para a mineração e posterior inserção automatizada nos metadados correspondentes no Tainacan (Figura 1). Cabe salientar que "Livro de registro" pode ser subdividido em: "Livro das Celebrações", "Livro das Formas de Expressão", "Livro dos Lugares" e "Livro dos saberes". Os demais dados foram fornecidos pelo Iphan e os "Documentos" foram categorizados para posterior importação (conforme item 4.2).

Inicialmente foram propostas 20 classificações para os documentos (a partir da análise qualitativa). Posteriormente, foi realizada a leitura automatizada dos textos, identificando-se padrões (no texto e no cabeçalho) utilizando regex2, visando à validação das categorias. Após esses processos, o Iphan realizou algumas adaptações e validou as categorias. Com isso, foram definidas 22 categorias de documentos (Tabela 1), baseadas na formatação e definição do seu tipo. Esse processo foi importante para a padronização dos nomes dos arquivos e identificação de documentos com metadados a serem preenchidos nos itens no Tainacan.

2 Regex é uma expressão regular flexível de identificação de cadeias de caracteres de interesse

Item	Categoria documental	Qde.
01	Parecer Técnico de Registro	98
02	Anuência para registro de bem cultural	76
03	Aviso no Diário Oficial da União	73
04	Pedido para registro de bem cultural	62
05	Ata de Reunião do Conselho Consultivo do Patrimônio Cultural (Registro)	54
06	Certidão de registro de bem cultural	52
07	Dossiê de Registro	52
08	Parecer do Conselho Consultivo	51
09	Titulação de registro de bem cultural	40
10	Plano de Salvaguarda do Bem Cultural Registrado	22
11	Estatuto Social	14
12	Termo de Averbação de registro de bem cultural	11
13	Extrato de Decisão da Câmara Técnica do Patrimônio Imaterial	10
14	Nota Técnica para Revalidação	6
15	Parecer Técnico de Reavaliação	5
16	Ata de Reunião do Conselho Consultivo do Patrimônio Cultural (Reavaliação)	2
17	Certidão e titulação de registro de bem cultural	2
18	Ata de Reunião da Câmara de Patrimônio Imaterial (Registro)	2
19	Ata de Reunião da Câmara de Patrimônio Imaterial (Revalidação)	2
20	Diretrizes para Instrução Técnica de registro de bem cultural	1
21	Anuência e pedido para registro de bem cultural	1
22	Plano de Salvaguarda do Bem Cultural Registrado	1

Tabela 1 - Categorias documentais definidas para renomear e classificar os documentos

Fonte: dados da pesquisa (2023)

A renomeação foi realizada considerando o formato do arquivo a ser renomeado (texto, fotografia e vídeo). No caso de fotografia e vídeo, foi definida a estrutura "Formato do arquivo" - "Título do bem cultural". Em relação aos documentos textuais foram adotadas as categorias documentais (definidas no item 4.2) seguindo-se a estrutura: "Categoria documental" - "Título do Bem Cultural" (Quadro 1). A renomeação dos arquivos foi executada, computacionalmente, utilizando-se um script em Python.

Nome Original	Categoria documental
Certidão.pdf	Certidão do registro de bem cultural - Arte Kusiwa.pdf
Aviso DOU (1).pdf	Aviso no Diário Oficial da União - Cajuina.pdf
Volume 3 - SÍTIO INVENTARIADO.pdf	Dossiê de Registro - Cavalo Marinho.pdf
FD 353-354 - Caboclinhos - Pernambuco - 2012 - Foto Felipe Peres Calheiros- Acervo CNFCP-IPHAN.jpg	Fotografia - Caboclinho.jpg

Quadro 1 - Exemplo da nomenclatura dos arquivos originais e as respectivas categorias para as quais foram renomeadas

Fonte: dados da pesquisa (2023)

3.3 Conversão dos documentos em imagem e em texto

O processo de conversão de documentos foi necessário para possibilitar a visualização no Tainacan e realizar a extração de dados textuais. Como muitos documentos estavam digitalizados, não foi possível realizar o OCR. Por isso, decidiu-se primeiramente converter os arquivos em imagem. Esse processo de conversão dos arquivos para imagem foi realizado apenas com o objetivo de extração dos dados de texto, não sendo tais imagens salvas e nem carregadas no Tainacan. Posteriormente os arquivos foram salvos em pdf, formato definido pelo Iphan para o armazenamento dos documentos.

O processo contou com a conversão das imagens, utilizando um script do python com a função pdf2image da biblioteca poppler-utils, e extração do texto utilizando o Tesseract (um software Open Source para OCR), configurado para receber textos em língua portuguesa.

3.4 Mineração de texto

Para a mineração dos textos dos documentos foram desenvolvidos programas em linguagem de programação Python, utilizando regex a fim de encontrar padrões no texto e extrair strings. Os dados dos documentos definidos (item 4.1) foram minerados para a carga automatizada nos respectivos metadados no Tainacan. Esse padrão foi criado para a mineração dos segmentos de texto relacionados a: "Livro de Registro", "Bem cultural", "Descrição" e "Data de Registro" (Figura 1).

Com os processos de categorização (item 4.2) e renomeação dos títulos dos arquivos de texto e mídias (item 4.3), também foi possível realizar a extração dos dados automaticamente. Além disso, foi possível a extração

dos dados de Autoria e Ano de criação dos documentos de mídia (fotografias e vídeos). Para a mineração dos dados de autores foi utilizado um modelo pré-treinado da biblioteca Spacy (pt_core_news_lg) e regex para o Ano de sua criação.

3.5 Carga de documentos

Os dados dos documentos e mídias mineradas foram inseridos em um csv padronizado e foi realizada a sua importação utilizando uma funcionalidade do próprio Tainacan. Cada linha no csv correspondeu a um item e cada coluna correspondeu a um metadado do tainacan. Os dados de cada metadado foram preenchidos com os valores obtidos pelo processamento de extração dos documentos ou fornecidos pelo próprio Iphan3. Nessa etapa foi realizado o processo de limpeza, organização e remoção de caracteres indesejados, utilizando-se do regex.

Foi realizada a carga automatizada dos 52 itens da coleção "Bens Culturais Imateriais Registrados". Cada arquivo de mídia foi previamente relacionado, via identificador único, ao seu bem cultural antes de ser importado em sua respectiva coleção. Esses identificadores foram utilizados para vincular os documentos e coleções criando-se hiperlinks entre os diversos itens de uma coleção, proporcionando a conexão e interatividade entre o bem cultural e seus documentos e mídias.

3.6 O acervo Legado do INRC

O acervo legado do INRC foi disponibilizado pelo Iphan, em pastas do Google Drive. O acervo corresponde a 9825 documentos textuais, 41428 fotografias, 3484 vídeos e 614 áudios4. Dessa forma, 188 projetos de identificação que formam o acervo foram automaticamente inseridos no Tainacan com a importação em massa.

Na primeira etapa foi realizada uma análise documental e identificação de metadados por meio da análise qualitativa visando à identificação de padrões passíveis de extração automatizada. Os metadados identificados foram incluídos no Tainacan para posterior importação dos dados, conforme relatório da meta 2.

A segunda etapa consistiu na categorização dos documentos por meio da análise dos conteúdos e cabeçalhos. Tal etapa foi necessária em razão da grande variedade terminológica dos nomes dos arquivos. Foram identificados utilizando expressões regulares padrões de 18 tipos de documentos diferentes, divididos em Fichas, Questionários e Anexos. Os itens não identificados nesses padrões são documentos de cunho cultural fora do padrão do INRC, que serão validados e categorizados futuramente utilizando outros métodos computacionais e manualmente pelos especialistas do Iphan e do Ibict.

3 https://docs.google.com/spreadsheets/d/1y8HhxnUMvyr7zSO4RIK08CfoYppvq6TD/edit?usp=sharing&oui-d=117398363879732217681&rtpof=true&sd=true

4 [3] Segundo levantamento realizado, disponível no Relatório de Meta 1. Posteriormente foram enviados alguns projetos, produto de mapeamentos, que compõem o acervo legado. Necessário realizar um levantamento futuro, procedimento realizado nos outros projetos do acervo.

Na terceira etapa realizou-se a padronização do título dos arquivos que consistiu na padronização e renomeação automatizada dos títulos dos arquivos (realizada a partir das categorias desenvolvidas na etapa 2). Importa ressaltar que os títulos foram estabelecidos de forma provisória, indicando mais estudos para definições futuras em um processo de curadoria.

A quarta etapa consistiu em um processo de transformação de todos os arquivos para o padrão PDF e a conversão de documentos digitalizados para um formato de texto legível pela máquina, por meio do OCR, permitindo a identificação e busca de texto no documento para pesquisa. Essa etapa foi necessária pelo fato de existirem arquivos digitalizados em diversos formatos, inviabilizando a extração de dados, de modo que não foi utilizado OCR como método de extração dos dados pela pouca existência de documentos digitalizados.

Na quinta etapa foi realizada a mineração de texto, durante a qual extraiu-se dos documentos um script criado na linguagem python com bibliotecas de leituras de texto em PDF. Os dados foram, então, salvos em um arquivo no formato csv no modelo aceito para importação pelo Tainacan.

A sexta e última etapa contemplou a importação dos documentos e mídias na aba de mídias do wordpress e a importação e visualização dos dados no software Tainacan, relacionando os itens criados no Tainacan com seus documentos, mídias e coleções semelhantes. Tal etapa foi realizada por meio da ferramenta de importação, disponível no Tainacan, a qual possibilitou a criação de itens em massa de diferentes coleções, a criação automática de uma página do item que exibe seus dados, documentos e mídias associadas.

O relacionamento dos itens foi realizado utilizando metadados de "relacionamento" criados na coleção do Tainacan (para itens em diferentes coleções, p. ex.: documento e projeto de identificação) e o campo special-dcument do csv a ser enviado (para deixar documentos e/ou mídias em anexo).

3.7 Alterações específicas de Coleções

3.7.1 Bibliografia

Durante o processo de estruturação da coleção bibliográfica foram concebidos e testados alguns modelos visando à organização, reutilização e visualização. O primeiro foi concebido em uma estrutura de seções elaborada de acordo com a classificação da bibliografia, contando com metadados específicos. O modelo proposto se baseou na norma ABNT 6023, porém, foi simplificado, pois os preenchedores das bibliografias não são especialistas. Caso contrário – ou seja, se tivéssemos proposto algo mais complexo –, entendemos que poderia haver desestímulo para o preenchimento das bibliografias pelos usuários.

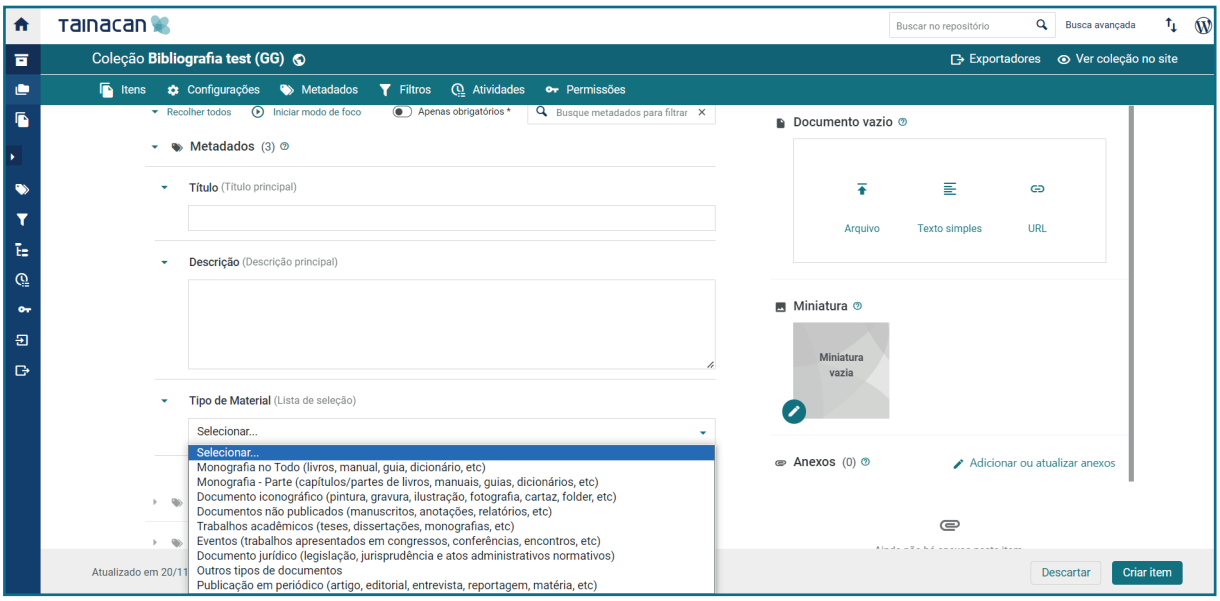


Figura 2 - Tela de metadados da coleção Bibliografia test

Fonte: Elaborado a partir da instalação do INRC (2023)

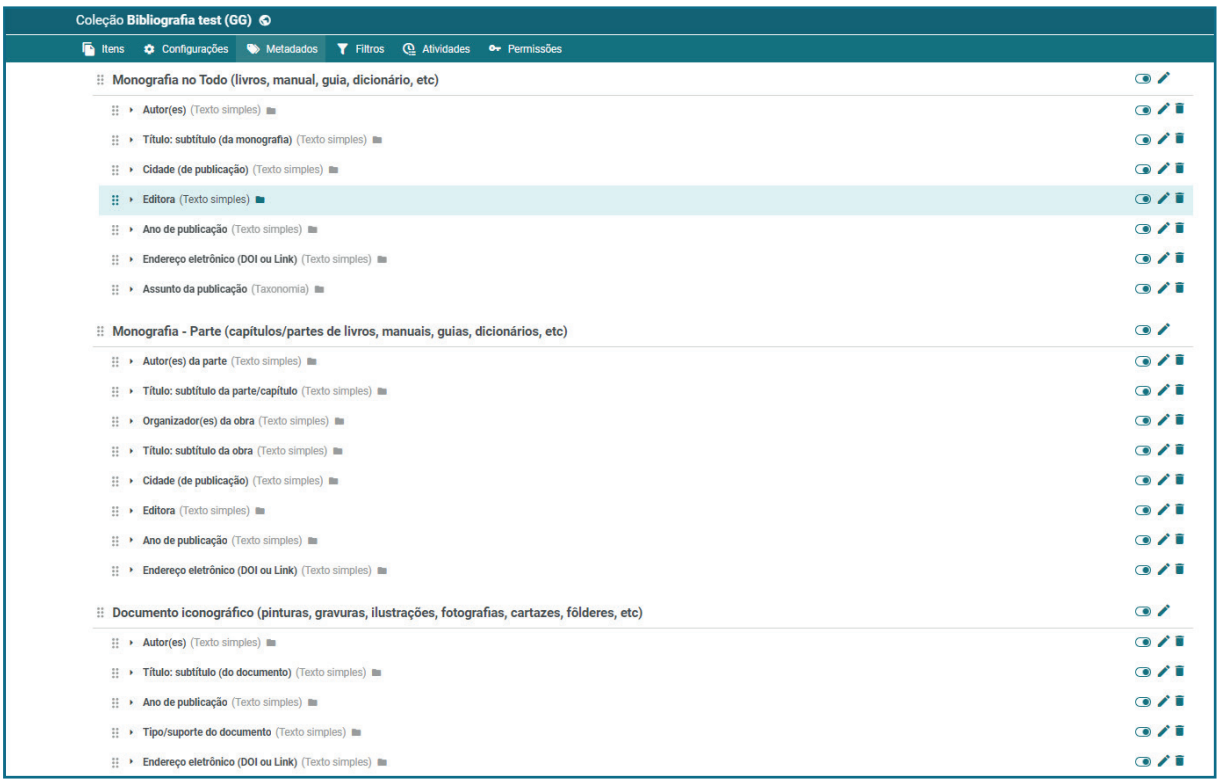


Figura 3 - Tela de metadados da coleção Bibliografia test

Fonte: Elaborado a partir da instalação do INRC (2023)

Neste protótipo existiriam metadados específicos para cada tipo de bibliografia, enriquecendo a qualidade de informação. Apesar dos benefícios, tal formato foi desconsiderado por não contemplar um fator muito importante: a impossibilidade de exportação para padrões internacionais como o Dublin Core. O maior receio da equipe foi a dificuldade no futuro quando o repositório estivesse "povoado" e fosse preciso exportá-lo pelas várias seções,

o que poderia acarretar na perda de informação ou na desestruturação de toda a coleção. Assim, esse tipo de modelagem foi descartado.

Com o estudo anterior de organização da coleção, foi possível propor uma nova modelagem para a coleção. Tomando como base as tipologias sugeridas, foi criado um metadado de lista de seleção contendo todas as tipologias bibliográficas. Tal configuração foi projetada, de forma geral, para induzir o usuário a enquadrar o tipo de bibliografia sem que houvesse conhecimentos de normas para elaboração de referências.

Sugere-se também a criação de uma taxonomia específica para os anos de publicação em vez de ser um metadado de texto simples, para que corra o melhor refino da busca.

3.7.2 Documentos

Na coleção de documentos foi realizada uma revisão para constar não apenas a autoria do documento, mas pessoas da equipe que participaram na construção do documento. Com isso, foi adicionado os metadados: "Contribuidor" e "Editor".

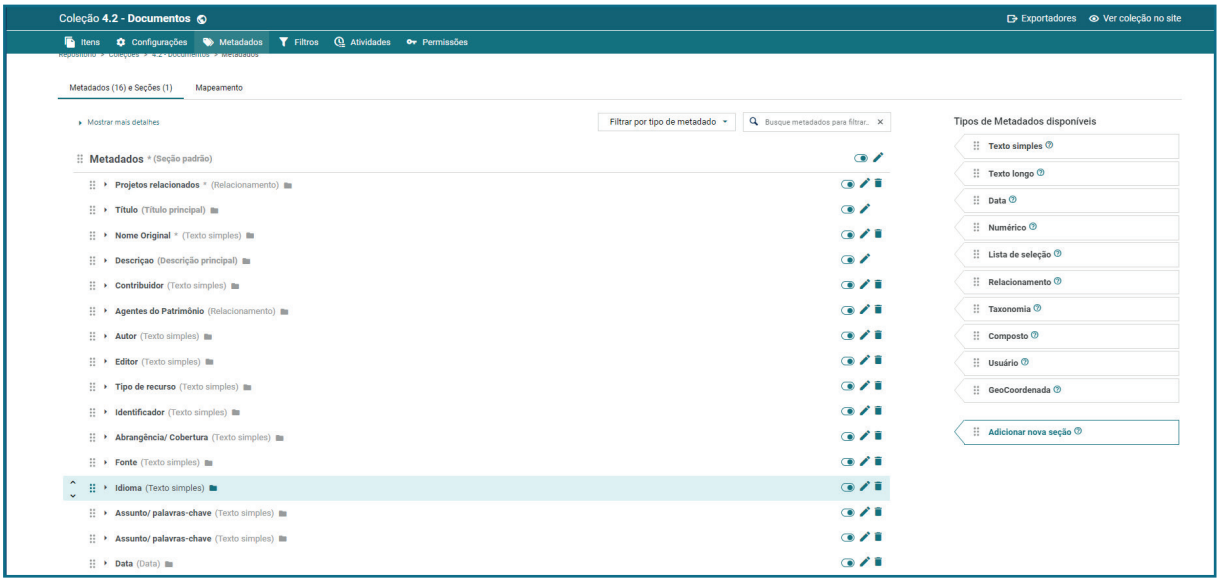


Figura 4 - Tela de metadados da coleção Documentos

Fonte: Elaborado a partir da instalação do INRC (2023)

Também foi adicionado um metadado de relacionamento que linka com a coleção de "Agentes do patrimônio" para manter as informações dos detentores.

3.7.3 Mídias - documento

A utilização de metadados mínimos, conforme definido pelas legislações de descrições arquivísticas e bibliográficas, é fundamental para referenciar mídias. Os metadados fornecem informações essenciais para identificar e recuperar registros específicos. Ao utilizar metadados mínimos, garante-se que há informações suficientes para distinguir e localizar os recursos, mídias ou documentos relevantes. A consistência

dos padrões de acordo com as legislações garante a interoperabilidade com diferentes acervos e tipos de acesso. Metadados estruturados são essenciais também para a gestão da preservação digital. Informações sobre a proveniência, o formato e as condições de acesso são cruciais para garantir a preservação a longo prazo dos bens culturais digitais.

Além disso, a descrição de documentos e mídias pode ajudar a prevenir a perda de informação. Os metadados fornecem informações sobre a origem, a criação e o conteúdo dos documentos, o que pode ajudar a protegê-los de danos e perdas.

Para o projeto de transformação digital do IRNC foram utilizadas legislações arquivísticas e padrões bibliográficos no sentido de auxiliar a seleção dos metadados que indicarão as mídias para organização, recuperação e disseminação da informação. Algumas legislações utilizadas foram:

3.7.4 Arquivísticas

- A Lei Federal nº 8.159/91, que instituiu a Lei de Arquivos, estabelece a obrigatoriedade da descrição arquivística para os documentos públicos e privados. A descrição arquivística é o processo de identificação, análise, organização e descrição de documentos, com a finalidade de torná-los inteligíveis e acessíveis.

- O Decreto nº 4.073/02, que regulamenta a Lei de Arquivos, determina que a descrição arquivística deve conter, no mínimo, as seguintes informações:

    - Identificação do documento: título, data, autor, número, tipo de documento;
    - Descrição do conteúdo: resumo do conteúdo do documento, principais temas abordados;
    - Contexto de criação: informações sobre a origem, a produção e o uso do documento;
    - Condições de uso: informações sobre a preservação e o acesso ao documento;

- O Decreto 10.278/2020, que dispõe sobre a digitalização de documentos públicos para fins de preservação e estabelece requisitos mínimos para a descrição de documentos digitalizados. Tais requisitos incluem:

    - Identificação do documento: título, data, autor, número, tipo de documento;
    - Descrição do conteúdo: resumo do conteúdo do documento, principais temas abordados;
    - Informações técnicas: resolução, formato;
    - Assinatura digital: assinatura eletrônica que garante a integridade e a autenticidade do documento.

3.7.5 Legislações Bibliográficas

- A Lei Federal nº 11.890/08, que instituiu o Sistema Nacional de Bibliotecas Públicas e, em seu artigo 10, estabelece que as bibliotecas públicas devem organizar e manter seus acervos de forma a facilitar o seu acesso e uso.

- A Resolução nº 14/2011 do Conselho Federal de Biblioteconomia (CFB), que estabelece as normas para a catalogação de documentos e determina que os metadados devem ser completos e precisos, de forma a permitir a recuperação e a identificação dos documentos.

- A norma internacional ISO 2709, que define o formato para intercâmbio de registros bibliográficos, é utilizada para a troca de metadados entre bibliotecas e outros sistemas de informação.

Para o projeto o esquema de metadados adotado foi o seguinte:

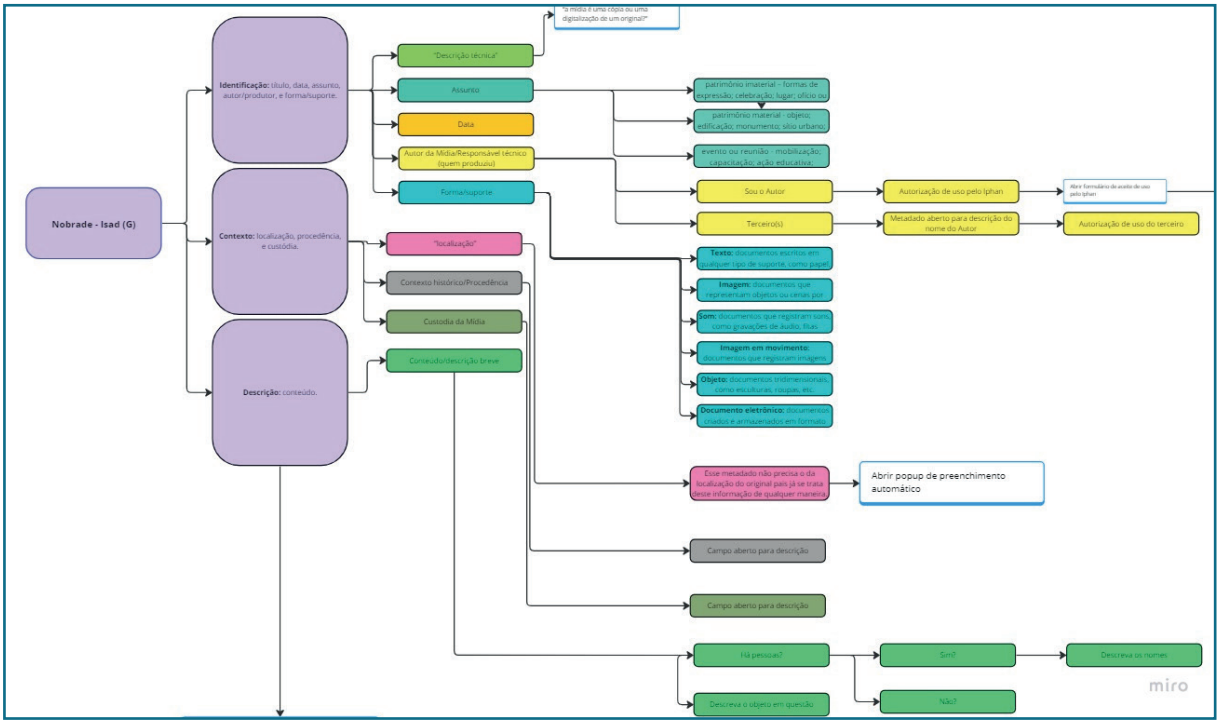


Figura 5 - Modelo de esquema de metadados para as mídias do IRNC

Fonte: Elaborado pelos autores (2023)

3.7.6 Mídias extraídas de documentos

As mídias foram coletadas por meio da biblioteca Fitz/PymuPDF para identificar imagens, tamanho e localização no arquivo, de modo que são identificados e selecionados aqueles com no mínimo 500 pixels de largura para extração e download. Os metadados das mídias não puderam ser coletados porque não estão estruturados e preparados para uma coleta simplificada. Recomendamos estudos aprofundados para identificação dos valores a serem considerados.

3.8 Carga do INRC

Dentro do Universo de 133 Projetos do INRC Legado disponibilizados pelo drive, cerca de 108 foram automaticamente inseridos no Sistema Tainacan com a importação em massa.

Vale salientar que, dos 133 Projetos do INRC Legado, cerca de 25 se encontravam compactados (formato .zip) e com um volume expressivo de dados para ser descompactado, alguns contendo cerca de 80gb.

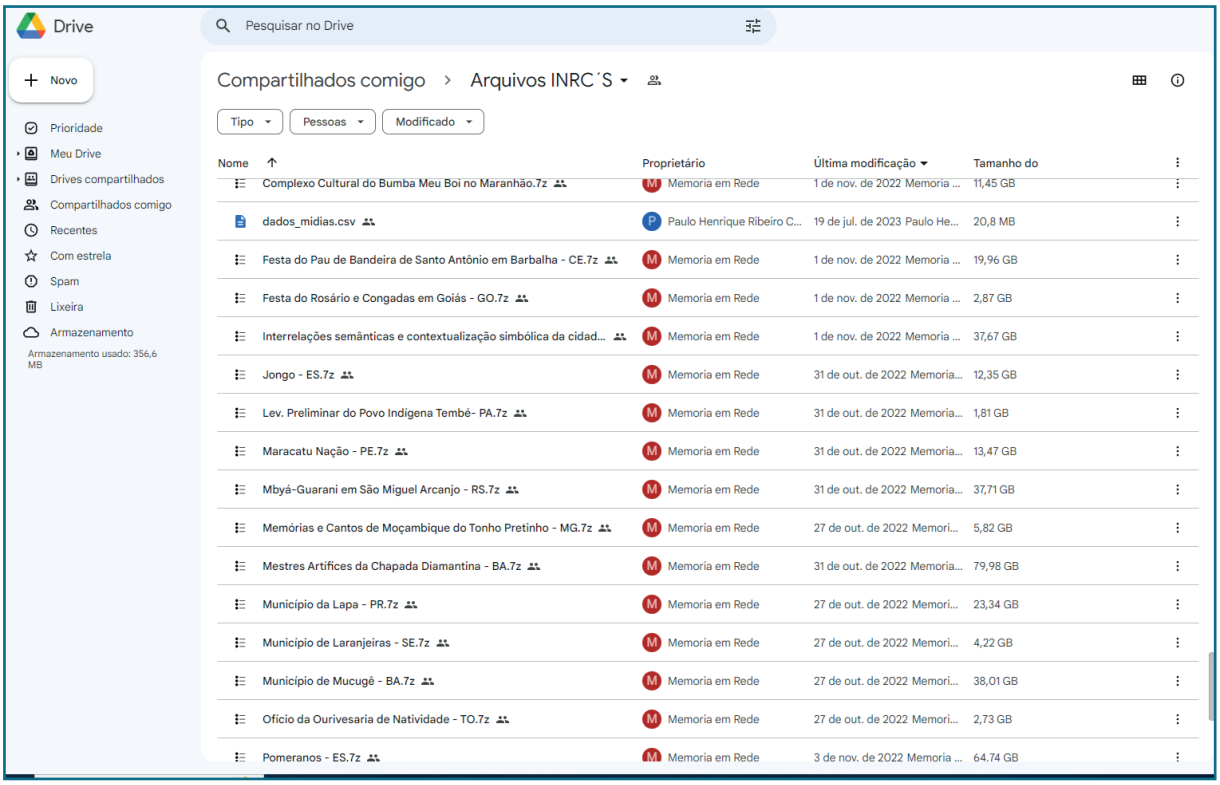


Figura 6 - Pasta compartilhada com os arquivos do Projeto

Fonte: Drive do INRC (2023)

Dessa forma, os projetos compactados não foram inseridos no Tainacan.

- **Dados inseridos na plataforma Tainacan -INRC**

  » Por meio da consulta à instalação Tainacan - INRC5, atualmente mantida pelo Ibict, foi possível realizar o seguinte diagnóstico6 acerca dos dados e documentos inseridos por meios dos processos automáticos supracitados:

5 Disponível em: http://colaborativo.ibict.br/tainacan-iphan-inrc

6 Link para documentos, mídias extraídas, planilhas com os relacionamentos: https://drive.google.com/drive/folders/1kY7MK-6DueE8a_QJd-2Q3lb7rY8OoMIlp?usp=sharing.

- **Coleção 1.1 Agentes do Patrimônio**

  - 1171 itens inseridos de todas as coleções
  - Metadados preenchidos: Nome do Agente; Projeto Relacionado; Tipo do Agentes
  - Fonte dos dados: Fichas e questionários com campo de Técnicos Responsáveis
  - Observação: Os dados devem ser verificados, limpos, normalizados e enriquecidos.
  - 241 itens inseridos - Exclusivamente da Arte Santeira e Frevo (verificar rascunhos e publicados).
  - Metadados preenchidos: Projeto relacionado; 1-Nome do Agente; 2 - Histórico da atuação cultural/patrimonial
  - Fonte dos dados: a partir da ficha Contatos.

- **Coleção 1.2 Organizações**

  - Observação: Nenhum item - revisar as fichas para sugerir a coleta dos dados.

- **Coleção 1.3 Proposta de identificação:**

  - Observação: Nenhum item - para essa coleção não existem dados legados.

- **Coleção 2.1 Projetos de identificação:**

  - Quantidade: 187 projetos diferentes
  - Metadados preenchidos: Região; Estado; Nome do projeto; Instituição Executora; Data de início; Data de Fim; Observação sobre o inventário; Pendências.
  - Fontes de informação: planilhas fornecidas pela equipe Iphan:
  - Tabela foco do inventário.xlsx
  - Lista dos inventários.xlsx
  - Observação: Verificar, limpar e normalizar os dados.
  - Consta os projetos Chimarrão, Festa do Nosso Senhor do Bonfim, mas estão sem documentação.

- **Coleção 2.2 Detalhamento de Projetos:**

  - » Observação: Nenhum item - para essa coleção ainda não foram designados dados legados.

- **Coleção 3.1 Identificação - Comunidade:**

  - » Observação: Nenhum item - para essa coleção ainda não foram designados dados legados.

- **Coleção 3.2 Identificação - Território**

  - » 558 itens (verificar rascunhos e publicados)

  - » Metadados preenchidos: 1-Nome principal do território; 1.2- Localização; 1.3 Descrição e relevância do território; Projetos relacionados; 4 - Paisagem e meio ambiente.

  - » Fontes de informação: Ficha de Localidade(F11); Ficha de Sítio (F10)

  - » Observação: Os documentos utilizados para a extração de dados que criaram o item foram relacionados, porém necessitam de validação. Verificar dados e duplicidades, normalizar e enriquecer.

- **Coleção 3.3 Identificação - Bem Cultural**

  - » 963 itens (verificar rascunhos e publicados)

  - » Metadados preenchidos: Projeto relacionado; 1.1 - Nome do bem cultural; 1.2 - Outros nomes do bem cultural; 1.3 Categoria de bem cultural; 1.3 Categoria de bem cultural; 1.4 - Localização; 1.5 - Descrição;

  - » Fontes de informação: Ficha de Identificação de bens culturais. O anexo Bens Culturais não foi utilizado. Esse documento deve ser explorado no futuro.

  - » Curadoria das mídias para representar de forma gráfica os bens culturais.

  - » O trabalho de seleção foi realizado em conjunto com a equipe do Iphan (Pedro e Sara) para poder garantir que as mídias estavam sob guarda dos direitos autorais pelo Iphan para divulgação. A seleção partiu primariamente das mídias com direitos resguardados, e por fim uma análise qualitativa de qual foto poderia corresponder melhor a um Bem cultural, de modo que nenhuma pessoa ficasse em evidência e sim no bem cultural específico.

  - » Observação: Os documentos utilizados para a extração de dados que criaram o item foram relacionados, porém necessitam de validação. Verificar dados e duplicidades, normalizar e enriquecer.

- **Coleção 3.4 Identificação - Grupos**

  » 51 itens publicados - Arte Santeira e Frevo

  » Metadados preenchidos: Projeto de identificação relacionado; 1 - Nome do Grupo; 1.1 - Apresentação; 2. Localização.

  » Fontes de informação: Ficha de Identificação de bens culturais. Foi realizada uma curadoria de dados pelo Pedro/Iphan para os casos da Arte Santeira e do Frevo.

  » Observação: Os documentos utilizados para a extração de dados que criaram o item foram relacionados, porém necessitam de validação. Verificar dados, normalizar e enriquecer.

- **Coleção 3.5 Identificação - Detentores**

  » 997 itens publicados

  » Metadados: Projetos relacionados; 1 - Nome completo; 1.1 - Nome como é conhecido no grupo ou comunidade; 1.4 - História de vida; - Localização; 4 - Aprendizado do bem; 2 - Localização; 4 - Aprendizado do bem; 4.2 - Transmissão de saberes.

  » Fontes de informação: Questionários.

  » Observação: Os documentos utilizados para a extração de dados que criaram o item foram relacionados, porém necessitam de validação. Verificar dados e duplicidades, normalizar e enriquecer.

- **Coleção 3.6 Diagnóstico comunitário**

  » Observação: Nenhum item - para essa coleção não existem dados legados.

- **Coleção 4.1 Bibliografia**

  » Modelo Atual (Coleção: Referências/Bibliografias)

  » 20206 itens publicados

  » Metadados preenchidos: Título, Bem Cultural Relacionado, Assunto Livre, Ano de Publicação.

  » Fontes de informação: Anexo Bibliografia (F1-A1)

  » Observação: Verificar dados, normalizar e enriquecer.

- **Coleção 4.2 Documentos**

  » 8985 itens

  » Metadados preenchidos: Projetos relacionados; Título; Nome Original; Agentes do Patrimônio; Autor; Editor.

  » Fontes de informação: Todas fichas e anexos originais foram adicionados como item nessa coleção.

  » Observação: Para preencher o metadado "nome original do documento" foi utilizado o título do arquivo digital fornecido pelo Iphan. A descrição do documento foi elaborada pelo Iphan. O título foi estabelecido em caráter provisório. Recomenda-se a realização de estudos para novas propostas. Necessário verificar relações.

- **Coleção 4. 3 Mídias**

  » 647 itens - Arte Santeira e Frevo

  » Metadados preenchidos: Projetos relacionados; Equipe (retirar registro); Título; Descrição Técnica; Assunto;Classificação; Localização da mídia original. Autoria própria

  » Fontes de informação: As mídias dos projetos Arte Santeira e Frevo passaram por um processo de verificação do documento de licença. Conforme explicitado na coleção de bem cultural (3.3-Identificação - Bem Cultural), com a subida das mídias foi necessário realizar uma seleção dos itens. A principal preocupação era a existência de mídias cujo o Iphan não tinha posse do direito de divulgação, e algumas mídias se encontravam nessa situação porque muitas pesquisas foram realizadas antes dessa política bem definida de direitos autorais e de divulgação em meio digital. Assim, as mídias que se encontram na situação de falta de documentação estão na plataforma Tainacan de forma privada, ou seja, apenas os gestores do Iphan têm acesso ao acervo.

  » Obs.: As outras mídias das coleções precisam ser inseridas no wordpress, juntamente com as mídias extraídas dos documentos. Observar e avaliar a presença das mídias que foram inseridas também nos anexos (p. ex.: Celebração-Recife).

4. CONSIDERAÇÕES FINAIS

O presente relatório apresentou os resultados da pesquisa que teve como objetivo o estudo de soluções automáticas para o tratamento e carregamento de dados e documentos nas bases do BCR e do Novo INRC. A criação desses processos algoritmos se deu de forma concomitante a estudos que identificaram, de forma qualitativa, os caminhos que deveriam ser percorridos, mediante a análise da documentação, do sistema de informação de destino e da necessidade e contexto dos usuários e do domínio temático da informação.

Como resultados, o BCR se encontra atualmente em utilização pelas equipes do Iphan e disponível para consulta pública. Como a plataforma foi desenvolvida para fins de protótipo, recomendamos uma análise dos dados para identificação de necessidades de melhoria da qualidade deles. Também se torna necessária e urgente a realização e implementação de novo projeto de design gráfico para a plataforma. Até o presente momento7, a instalação ainda se encontra nos servidores do Ibict.

Uma versão do INRC, que está hospedada nos servidores do Iphan, encontra-se on-line e disponibiliza duas coleções (Arte Santeira e Frevo) que foram alvo de processos de curadoria para servirem de "projetos-modelo" para os demais. Os outros 186 projetos adicionados foram acompanhados somente de informações básicas e com o indicativo de que o repositório estava em processo de construção. Outra versão do INRC, que está hospedado nos servidores do Ibict, é o ambiente de produção para a inserção dos dados legados. A versão que está no Iphan é uma cópia, inicialmente pensada apenas para consulta, porém posteriormente, em caráter emergencial, a Coide disponibilizou o ambiente para a submissão de conteúdos de uma organização parceira. Assim, é recomendada a realização de uma análise prévia comparativa dos dados de ambas as instalações antes de qualquer procedimento de transferência e substituição de versões.

Como resultado dos trabalhos realizados, diversas informações foram identificadas na documentação legada, a qual estava nos moldes do antigo INRC, permitindo a compatibilização com a nova versão do INRC. Dessa forma, tanto foram criados itens novos como também a documentação original foi inserida de forma relacionada, a fim de permitir tanto a consulta das fontes dos dados como possibilitar uma futura análise qualitativa aprofundada da documentação, que busca permitir a identificação e extração para inserção de novos dados relevantes no sistema.

Vale ressaltar que, devido ao caráter de pesquisa experimental, buscou-se analisar a documentação e criar estratégias para a extração e inserção em massa dos dados a fim de otimizar o processo de representação das informações de bens culturais identificados pelo inventário. Embora tenhamos realizado avanços significativos, são necessários projetos futuros que deem continuidade aos desdobramentos da pesquisa, para que se possa gerar um serviço de informação eficiente, que forneça informação confiável e de qualidade, tanto no conteúdo

7 4 de janeiro de 2023.

quanto na forma que é apresentada. Os dados precisam ser verificados e validados, normalizados e enriquecidos, de modo que sua arquitetura também precisa ser reavaliada de acordo com novos desenvolvimentos do Tainacan e de novas necessidades identificadas. Assim sendo, é urgente a realização e implementação de novo projeto de design gráfico para a plataforma.

REFERÊNCIAS

ALONSO, Maria Teresa; RODRIGUES, Ana Cristina. **Metadados**: uma introdução. Rio de Janeiro: Editora FGV, 2011.

CORDEIRO, Maria do Rosário. **Metadados para bibliotecas**: fundamentos e prática. São Paulo: Editora Polis, 2016.

BRASIL. Conselho Federal de Biblioteconomia. **Resolução nº 14/2011, de 23 de fevereiro de 2011**. Aprova as normas para a catalogação de documentos.

BRASIL. **Decreto nº 10.278, de 20 de julho de 2020**. Dispõe sobre a digitalização de documentos públicos para fins de preservação. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/decreto/d10278.htm. Acesso em: 4 jan. 2023.

BRASIL. **Decreto nº 4.073/02, de 3 de janeiro de 2002**. Regulamenta a Lei nº 8.159, de 8 de janeiro de 1991, que dispõe sobre a política nacional de arquivos e dá outras providências. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto/2002/d4073.htm. Acesso em: 4 jan. 2024.

BRASIL. **Lei Federal nº 11.890/08, de 13 de novembro de 2008**. Institui o Sistema Nacional de Bibliotecas Públicas e dá outras providências. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11890.htm. Acesso em: 4 jan. 2024.

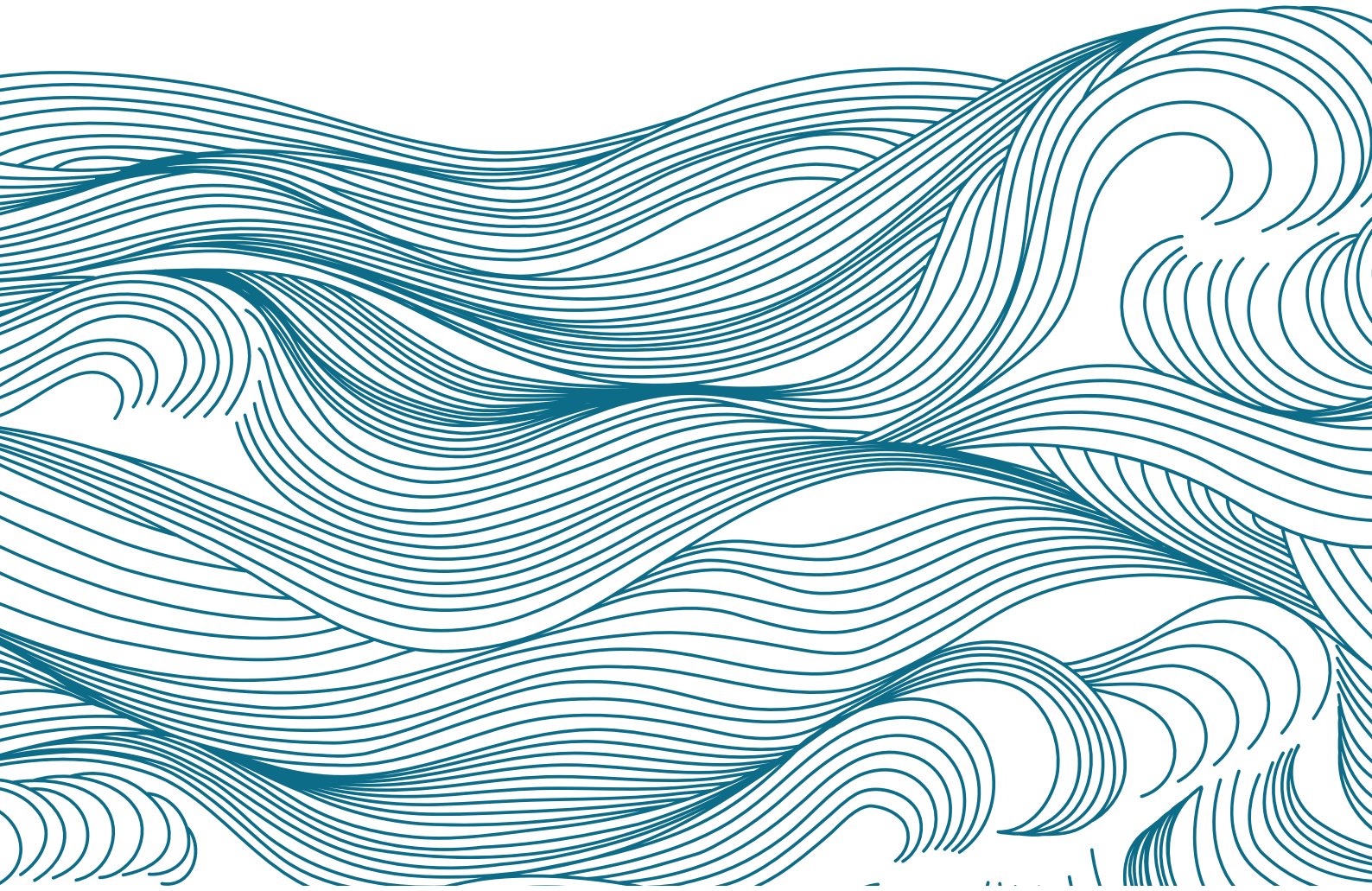
BRASIL. **Lei Federal nº 8.159/91, de 8 de janeiro de 1991**. Dispõe sobre a política nacional de arquivos e dá outras providências. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8159.htm. Acesso em: 4 jan. 2024.

HARNAD, Stevan *et al*. The Access/Impact Problem and the Green and Gold Roads to Open Access. **Serials Review**, [*s. l.*], v. 30, n. 4, p. 310-314, 2004. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0098791304001480. Acesso em: 4 jan. 2004.

SÁ, Maria do Carmo. **Metadados para documentos arquivísticos**: uma abordagem conceitual. Rio de Janeiro: Editora FGV, 2014.

SHINTAKU, M.; VIDOTTI, S. A. B. G. Bibliotecas e repositórios no processo de publicação digital. **BIBLOS**, [*s. l.*], v. 30, n. 1, p. 61–80, 2016. Disponível em: https://periodicos.furg.br/biblos/article/view/5762. Acesso em: 4 jan. 2024.

WEITZEL, S. da R. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 51–71, 2006. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/19. Acesso em: 4 jan. 2024.



ibict
Instituto Brasileiro de Informação em Ciência e Tecnologia

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO