Recuperando e Tornando Reusáveis e Interoperáveis Tesauros Brasileiros de Cultura e Ciência em Formato Textual

Recovering and Making Reusable and Interoperable Brazilian Thesauri of Culture and Science in Textual Format

Francis Bento Marques (1), Carlos Henrique Marcondes (2)

(1) Universidade Federal dos Vales do Jequitinhonha e Mucuri, Brasil, fbmarques@gmail.com

(2) Universidade Federal Fluminense, Brasil, ch_marcondes@id.uff.br

Resumo

Vários tesauros brasileiros nas áreas de Cultura e Ciência tornaram-se "órfãos" ou "zumbis" devido à descontinuidade de projetos e ao uso de formatos não estruturados (como PDF) ou plataformas fechadas, o que impede seu reuso e interoperabilidade. Esta pesquisa propõe soluções tecnológicas para tornar esses vocabulários migráveis e integráveis a novos sistemas. A metodologia, de natureza exploratória e aplicada, utiliza o software TemaTres para converter formatos textuais em SKOS. O experimento inicial consistiu na importação da Tabela de Áreas do Conhecimento CAPES/CNPq e do Tesauro Brasileiro de Ciência da Informação. Os resultados confirmam a viabilidade de transformar esses Sistemas de Organização do Conhecimento (SOCs) em formatos processáveis por máquina, embora o processo revele desafios técnicos na extração de dados. O trabalho discute a relevância desses vocabulários para a Web Semântica e Big Data, abordando também questões de políticas públicas e direitos autorais. Com a exportação bem-sucedida para SKOS, os tesauros analisados tornaram-se efetivamente reusáveis, demonstrando um caminho para a preservação e evolução do patrimônio informacional e científico brasileiro.

Palavras-chave: Tesauro; Vocabulário; Cultura; Reuso; Interoperabilidade; SKOS; TemaTres

MARQUES, Francis Bento; MARCONDES, Carlos Henrique. Recuperando e Tornando Reusáveis e Interoperáveis Tesauros Brasileiros de Cultura e Ciência em Formato Textual. *Brazilian Journal of Information Science: research trends*, vol. 20, publicação contínua, 2026, e026003. DOI: https://doi.org/10.36311/1981-1640.2026.v20.e026003.

Abstract

Several Brazilian thesauri in fields such as Culture and Memory have become "orphans" or "zombies" following the discontinuation of their original projects. Often available only in unstructured formats (PDF) or closed platforms, these vocabularies face significant barriers to reuse and interoperability. This research proposes technological solutions to make such resources migratable and integrable into new systems. The methodology, exploratory and applied, utilizes the TemaTres software to convert textual formats into SKOS. The initial experiment involved importing the CAPES/CNPq Table of Knowledge Areas and the Brazilian Thesaurus of Information Science. Results confirm the feasibility of transforming these Knowledge Organization Systems (KOS) into machine-processable formats, despite technical challenges in data extraction. The study discusses the relevance of these vocabularies for the Semantic Web and Big Data, while also addressing public policy and copyright issues. With the successful export to SKOS, the analyzed thesauri are now fully reusable, demonstrating a viable path for the preservation and evolution of Brazilian informational and scientific heritage.

Keywords: Thesaurus; Vocabulary; Culture; Reuse; Interoperability; SKOS; TemaTres

1 Introdução

Vocabulários fornecem semântica (metadados) aos dados e a seus conteúdos. São um dos instrumentos mais importantes para a padronização de conteúdos de conjuntos de dados como os Dados Abertos CAPES, os do Portal Brasileiro de Dados Abertos [(1)], os do DATASUS os do IBGE ou de acervos digitais em Cultura, Memória e Patrimônio disponíveis na Web como a BN Digital, Brasiliana Iconográfica, Brasiliana Fotográfica, Brasiliana Museus (Marcondes; Souza, 2018), etc. Favorecem a interoperabilidade entre conjuntos de dados e, nas consultas pela Web, permitem que usuários remotos realizem consultas distribuídas e explorarem os conteúdos de um recurso, minimizando o efeito "caixa-preta".

O desenvolvimento de vocabulários é um processo lento, caro, de maturação a longo prazo. O desenvolvimento de vocabulários é uma tarefa complexa, que demanda tempo e recursos. O Getty Vocabulary Program dedicou quase três décadas para a criação de tesauros que podem ser usados como bases de conhecimento, ferramentas de catalogação e documentação e assistentes de pesquisa *on-line* (Baca, 2016, p. 20). Muitos projetos de vocabulários no Brasil foram paralisados ou foram encerrados, gerando os chamados vocabulários "órfãos" ou "zumbis" (NISO, 2017), referência a vocabulários cujos projetos foram descontinuados e, portanto, não são mais atualizados e mantidos. Isto se dá principalmente porque os projetos de vocabulários terminam e não há mais recursos para mantê-los.

MARQUES, Francis Bento; MARCONDES, Carlos Henrique. Recuperando e Tornando Reusáveis e Interoperáveis Tesauros Brasileiros de Cultura e Ciência em Formato Textual. *Brazilian Journal of Information Science: research trends*, vol. 20, publicação contínua, 2026, e026003. DOI: https://doi.org/10.36311/1981-1640.2026.v20.e026003.

Infelizmente este é um grande problema. Esforços e recursos durante vários anos foram gastos no desenvolvimento de tesauros no Brasil que acabam desperdiçados pelo fim dos projetos e recursos. Entre os mais conhecidos estão o THESAURUS para acervos museológicos (Ferrez; Bianchini, 1987) e a sua mais recente versão para a Web, o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros, o Tesauro do Folclore e Cultura Popular, os Cabeçalhos de Assunto da Rede Bibliodata (Oliveira; Alves; Vicente 1997), o THESAURUS de acervos científicos (Granato, 2010), "Resultado de um projeto de cooperação entre 14 instituições brasileiras e portuguesas, o projeto que originou o Thesaurus foi coordenado pelo Museu de Astronomia e Ciências Afins (MAST), no Brasil, e pelo Museu de Ciências da Universidade de Lisboa (MCUL), em Portugal" **(2)**, entre os mais conhecidos.

Existem ainda outros tesauros importantes para a ciência e cultura brasileira (Eletrobras, 1992) que foram resultado de projetos e consumiram recursos importantes, mas que hoje têm seu alcance limitado por não estarem em meio digital nem disponíveis na Web, como o TESAURO DE CULTURA MATERIAL DOS ÍNDIOS DO BRASIL (Motta, 2006), cujo desenvolvimento foi resultado de um projeto apoiado pela UNESCO (2022) e que só existe em versão textual em formato PDF.

Embora estes vocabulários sejam importantes para a Cultura e Ciência no Brasil e tenham grande potencial para serem reusados, expandidos, desenvolvidos, este reuso é difícil. Apesar de disponíveis em meio digital, têm seus conteúdos "prisioneiros" de formatos textuais não estruturados **(3)** cujo conteúdo só pode ser visualizado, mas não processado por máquina; são operados a partir de plataformas tecnológicas proprietárias ou obsoletas, se constituindo em verdadeiros "silos de informação" (Shahrokni; Söderberg 2015). O reuso destes vocabulários é praticamente impossível, permitindo, quanto muito, sua consulta em linha, como um tesauro impresso, mas não sua migração, extração, desenvolvimento e incorporação a outros sistemas de informação.

Como tornar reusáveis tesauros brasileiros que estão disponíveis em formato textual digital não estruturado ou que estão "prisioneiros" de plataformas tecnológias que, apesar de estarem

disponíveis na Web, não permitem a interoperabilidade, migração/exportação para outros sistemas e plataformas?

Este trabalho discute esta questão e propõe uma solução tecnológica usando o software de gestão de tesauros TemaTres para converter estes vocabulários para SKOS – Simple Knowledge Organization System –. O SKOS é, na prática, um formato que utiliza as tecnologias da Web Semântica e é avalizado pelo W3C como padrão para a interoperabilidade de SOCs – Sistemas de Organização do Conhecimento – plataformas ou aplicativos gerenciadores como o Tecer, o MultiTes, o PoolParty Thesaurus Manager e o próprio TemaTres. O trabalho relata um experimento inicial usando as facilidades de importação de SOCs em formato textual para o Tematres para, posteriormente, exportá-los para SKOS. Foram importados dois SOCs, a Tabela de Áreas do Conhecimento (TAC) CAPES/CNPq e o Tesauro Brasileiro de Ciência da Informação (TBCI).

O trabalho tem, portanto, como objetivo discutir possíveis soluções tecnológicas para que tesauros brasileiros que estão disponíveis em formato textual digital não estruturado ou que estão "prisioneiros" de plataformas tecnológias fechadas possam ser interoperáveis, migráveis, exportáveis para outros sistemas e plataformas, enfim, reusáveis.

O trabalho está estruturado como se segue. Após esta Introdução a seção 2 expõe a metodologia. A seção 3 discute o papel dos vocabulários brasileiros em Cultura e Ciência diante do "Big Data" e da Web Semântica. A seção 4 descreve o experimento desenvolvido, dividida em subseções específicas sobre a importação da TAC e a importação do TBCI. A seção 5 discute os resultados do experimento. Por fim, a seção 6 apresenta as considerações finais.

2 Metodologia

Esta é uma pesquisa em andamento, de natureza exploratória e aplicada. Os casos de tesauros "órfãos" ou "zumbis" no contexto brasileiro são bastante conhecidos na prática dos profissionais da área. Além dos já citados, que são os candidatos naturais para serem objetos deste

experimento, dada a sua relevância para a Ciência e Cultura brasileiras, possivelmente outros possam ser identificados na literatura.

Para identificar experiências prévias do uso do software TemaTres para importar tesauros em formato textual foi feita uma pesquisa bibliográfica na base BRAPCI. O uso do termo "tematres" recuperou 15 referências, a maioria delas sobre usos em projetos específicos ou comparação com outros softwares de tesauros. Como o nosso interesse era mais específico, utilizou-se então o seu formulário de busca booleana com a seguinte estratégia de busca:

`tematres importação text*`, usando a opção "Busca pela query".

Não foi recuperada nenhuma referência. Eliminou-se então o termo "text*" e utilizou-se então o Formulário de Termos Booleanos da interface da BRAPCI, como os seguintes termos:

`"tematres" AND "importação"`

Foram então recuperadas 2 referências, Santos et al. (2018a) e Santos et al. (2018b), que tratam da importação de arquivos textuais em PDF para o TemaTres, mas não forneceram maiores detalhes sobre como formatar o arquivo TXT para importação, sendo essa a questão chave para importar com sucesso tesauros para o TemaTres.

O próprio sítio web do Tematres também disponibiliza uma série de manuais e guias que serviram de material de consulta.

Para a proposta de solução tecnológica utilizaram-se as funcionalidades de importar tesauros em formato textual do Tematres. Foi realizado um experimento importando a partir de arquivos texto para o TemaTres a TAC e o TBCI. A TAC foi obtida da página https://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf em 25/10/2025 e o TBCI foi obtido da página http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/copy_of_TESAUROCOMPLETOFINALCOMCAPA24102014.pdf/at_download/file em 27/11/2025. Uma vez importados, o TemaTres permite então que estes SOCs sejam exportados em formato SKOS, que é o objetivo deste trabalho.

Para viabilizar essa transparência e permitir a continuidade de novos estudos, o código-fonte dos scripts Python, bem como os arquivos estruturados gerados durante o experimento, estão disponíveis publicamente no repositório GitHub (https://github.com/fbmarques/labvoc/).

Para a conversão dos dados, foi desenvolvida uma rotina em Python utilizando bibliotecas como re (para expressões regulares) e sqlite3. O processo iniciou-se com a conferência automatizada das quantidades de termos extraídos dos arquivos originais, assegurando que o volume de dados hierárquicos (como Grandes Áreas e Subáreas da Tabela CAPES ou os termos do Tesauro) correspondesse exatamente aos totais esperados nos documentos em PDF.

Após essa etapa inicial, foi realizada uma validação técnica através de consultas diretas ao banco de dados do Tematres para confirmar a integridade da importação. Como etapa final de controle de qualidade, executou-se uma busca manual por amostragem: selecionaram-se aleatoriamente termos nos PDFs originais para verificar se constavam corretamente no sistema digital, garantindo assim que a grafia e a estrutura hierárquica foram preservadas fielmente.

Detalhes técnicos deste experimento são descritos na seção 4.2 Uso da funcionalidade de importação de arquivos textuais para o TemaTres.

3 Big Data e vocabulários brasileiros para a Cultura e Ciência

A onipresença e ubiquidade das tecnologias da informação nas mais diferentes atividades humanas é uma realidade que não deixou de afetar também as ciências e a cultura. Esta situação tornou disponível uma grande quantidade de dados em formato digital, fenômeno conhecido como Big Data. Big Data tem sido usado para designar o fenômeno que descreve a enorme quantidade de dados digitais que estão sendo criados em enorme velocidade, grande heterogeneidade como resultado de atividades humanas, sociais, econômicas, científicas e culturais centradas na web.

No Brasil vivemos também grande disponibilidade de dados digitais, resultado da POLÍTICA DE DADOS ABERTOS.

Para que estes dados possam ser usados e tratados pelas TIs é necessário que sejam padronizados semanticamente e processáveis por máquinas, para que possam revelar padrões,

"insights" ou agrupamentos. Vocabulários enquanto instrumentos de padronização semântica, têm papel fundamental nesta questão. Vocabulários podem ser classificados em vocabulários ou esquemas de metadados, que definem as propriedades descritivas dos dados e vocabulários de valores padronizados para dados (Zeng, 2019, p. 127).

Por sua vez, acervos digitais brasileiros em Ciência e Cultura disponíveis na Web como a BN Digital, Brasiliana Iconográfica, Brasiliana Museus, Museu Histórico Nacional, Biblioteca Brasiliana Mindlin, Pinacoteca de São Paulo, Arquivo Nacional, CPDOC/FGV, entre outros, demandam instrumentos semânticos para facilitarem seu acesso e interoperabilidade. No contexto brasileiro já existem alguns destes instrumentos, como os já citados THESAURUS para acervos museológicos e a sua versão para a Web, o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros, o Tesauro do Folclore e Cultura Popular, o THESAURUS de acervos científicos, os Cabeçalhos de Assunto da Biblioteca Nacional referentes a Brasil – História **(4)**, os Cabeçalhos de Assunto da Rede Bibliodata, entre outros. Estes tesauros favorecem a interoperabilidade entre conjuntos de dados, viabilizam consultas transversais pela Web a vários acervos simultaneamente, permitem a usuários remotos explorar os conteúdos do recurso, minimizando o efeito "caixa-preta" (Marcondes; Souza, 2018).

A necessidade de vocabulários estruturantes, em temas fundamentais de interesse para a Cultura e Ciência, padronizados, mais genéricos, não exclusivos de um acervo, é cada vez uma condição essencial para disponibilizar e integrar acervos digitais na Web. Por exemplo, seria importante estruturar os cabeçalhos de assunto da BN sobre "Brasil- História" como um tesauro, desenvolver uma lista de lugares geográfico/históricos do Brasil **(5)**.

Infelizmente temos poucos vocabulários na área de Cultura e Ciência que não sejam exclusivos de uma instituição, foram pensados assim. Se ressentem de desenvolvimento e manutenção, são de uso restrito a um setor, acervo ou instituição, seus projetos foram descontinuados. Este é um dos maiores problemas dos vocabulários em Cultura e Ciência no Brasil.

Outro problema brasileiro é a profusão de vocabulários que se sobrepõem: existem diversos tesauros e vocabulários jurídicos **(6)**, existem vários sistemas de cabeçalhos de assunto nas

bibliotecas em geral e nas bibliotecas universitárias brasileiras em particular [7]. Há um retrabalho e duplicação de esforços por um lado, ao lado de áreas temáticas não são supridas por nenhum vocabulário ou tesauro.

4 Descrição do experimento

Soluções para a questão formulada são várias e se desdobram em vários aspectos. Apontaremos e discutiremos aqui principalmente as soluções técnicas, o experimento realizado com o software TemaTres e seus resultados. Como desdobramento das soluções técnicas discutiremos também na seção de Considerações finais e as relativas às políticas públicas que contemplem vocabulários no âmbito das políticas públicas brasileiras de Cultura.

4.1 SKOS como formato interoperável para Sistemas de Organização do Conhecimento

No que diz respeito às soluções técnicas, o SKOS (*Simple Knowledge Organization System Primer*) é um modelo de dados que se constitui, na prática, em um formato padronizado e interoperável. O SKOS explicita estes elementos estruturais de um SOC para que cada um deles possa ser processado por máquinas. O SKOS é padrão recomendado pelo W3C. É compatível com a proposta da Web Semântica, seu modelo de dados é representado em RDF – *Resource Description Framework* – (RDF 2004) e reproduz e torna acionável por máquina toda a estrutura de um tesauro conforme especificado na Norma ISO 25964: relações paradigmáticas, relações sintagmáticas, relações linguísticas Use e Usado por, além de Notas de Escopo.

A preocupação com a interoperabilidade em todas as áreas e aplicações é recorrente e crescente, motivada principalmente pelo surgimento da Web em fins do século XX e, mais recentemente, pelo projeto da Web Semântica (Berners-Lee; Hendler; Lassila, 2001). Na área de sistemas de organização do conhecimento a mais recente norma relativa a tesauros, a Norma ISO 25964 tem como focos exatamente a interoperabilidade entre tesauros e inclui também a modelagem de tesauros (Dextre Clarke; Zeng, 2012).

A interoperabilidade é contemplada ao prever relações de mapeamento entre termos de um tesauro para termos de outro. Além disso, esta Norma inclui um modelo conceitual de um tesauro

(Will, 2012). Neste estão previstas classes como conceitos, termos, rótulos linguísticos (labels), grupos de conceitos, entre outros, e relacionamentos entre estas classes, como os Termos Genéricos, Termos Específicos, Termos Relacionados, Use e Usado por. O SKOS reproduz, representando em RDF, estas classes e relacionamentos, permitindo assim que um tesauro, com todos os seus conceitos e relações entre eles seja interoperável, possa ser portado de um sistema para outro.

O SKOS permite exportar e importar vocabulários de um sistema para outro. Esta funcionalidade de exportar e importar vocabulários entre sistemas possibilita que um vocabulário possa ser incorporado, integrado e reusado por outros sistemas/bases de dados de gestão de acervos em Ciência e Cultura. Isso permitirá que o esforço despendido na criação de vocabulários possa ser facilmente reutilizado por outras instituições de Ciência e Cultura.

4.2 Uso da funcionalidade de importação de arquivos textuais para o TemaTres

O TemaTres aceita a importação de dois "layouts" diferentes de arquivos texto. O primeiro é o arquivo texto tabulado, no qual a hierarquia de uma taxonomia é dada pelas sucessivas tabulações de arquivo texto; um exemplo de arquivo texto com este "layout" é o que foi utilizado como entrada para importar a TAC, como é mostrado no fragmento a seguir:

```
60700009 Ciência da Informação
    60701005 Teoria da Informação
        60701013 Teoria Geral da Informação
        60701021 Processos da Comunicação
        60701030 Representação da Informação
```

O segundo "layout" é o arquivo texto etiquetado, com as relações entre termos de um SOC assinaladas por etiquetas; um exemplo de arquivo texto com este "layout" é o que foi utilizado como entrada para importar o TBCI, como mostrado no fragmento a seguir:

índices de periódicos
ING: periodical indexes (UF journal indexes)
BT: índices
RT: periódicos
RT: periódicos científicos

RT: serviços de indexação e resumo
RT: serviços de informação
BT: 2.1.1 Representação da informação
BT: 7.1 Tipos de Documento

Em arquivos texto etiquetados o TemaTres reconhece as seguintes etiquetas.

BT: Termo Genérico (Broader Term).
NT: Termo Específico (Narrower Term).
RT: Termo Relacionado (Related Term).
UF: Usado Para (Used For).
USE: Termo preferencial a ser usado.
SN: Nota de Alcance (Scope Note).

4.2.1 Importação de texto e conversão para SKOS da TAC

O processo de importação da Tabela de Áreas do Conhecimento (TAC) iniciou-se com a extração de texto do documento PDF oficial disponibilizado pela CAPES. Foi desenvolvido um "script"em linguagem Python utilizando a biblioteca PyPDF2 para converter o conteúdo do PDF para texto puro. Esta etapa inicial, embora crucial, produziu um resultado bruto que mantinha apenas parcialmente a estrutura de indentação original, mas perdia as complexas relações hierárquicas e introduzia numerosos ruídos, como quebras de página e inconsistências de formatação. O resultado foi um arquivo de texto que necessitava de um tratamento significativo antes de poder ser importado.

A análise do texto extraído revelou a fragilidade da conversão direta, um problema comum ao lidar com o formato PDF, que é otimizado para visualização e não para extração de dados. Foram identificados múltiplos problemas, como a fragmentação de termos, a perda de códigos numéricos que definem a hierarquia, e a junção incorreta de linhas de texto. Relatórios gerados por scripts auxiliares, que comparavam a estrutura extraída com a estrutura visual do documento, apontaram centenas de "campos fora do lugar" e "campos vazios". Essas inconsistências tornavam o arquivo textual inadequado para uma importação direta em sistemas de gerenciamento de tesauros como o TemaTres, que exigem uma estrutura lógica e consistente.

Para superar esses desafios, em vez de uma única solução automatizada, foi desenvolvido um conjunto de scripts em Python para realizar um pré-processamento iterativo. Este processo incluiu a remoção de caracteres de controle, a normalização de codificação de texto para UTF-8,

e a aplicação de regras heurísticas para reconstruir a hierarquia original baseada nos padrões de indentação e nos códigos numéricos. O resultado deste esforço foi a criação de um arquivo "Markdown" (tabela_areas_conhecimento.md), uma versão limpa e estruturada do tesauro, onde a hierarquia hierárquica foi fielmente representada através da indentação (tabulações), servindo como uma "fonte da verdade" para as etapas subsequentes.

De posse do arquivo "markdown" devidamente estruturado, a importação para o TemaTres foi realizada com sucesso. O software foi configurado para interpretar o layout de texto tabulado (indentado), onde cada nível de indentação corresponde a um nível na árvore hierárquica do tesauro. O arquivo foi processado pelo TemaTres, que mapeou corretamente os termos e suas relações de "Termo Genérico" (BT) e "Termo Específico" (NT). O resultado final foi a recriação completa da TAC dentro do TemaTres como um vocabulário navegável, consistente e pronto para ser exportado no formato SKOS, cumprindo o objetivo de transformar um documento estático em um recurso de conhecimento interoperável

4.2.2 Importação de texto e conversão para SKOS do TBCI

Para o experimento foi desenvolvida uma metodologia de extração e estruturação automatizada de dados a partir do arquivo original em formato PDF/Texto ("Tesauro Brasileiro de Ciência da Informação"). O processo foi realizado por meio de scripts personalizados em linguagem Python, divididos nas seguintes etapas:

- **Extração e Pré-processamento**

O conteúdo do arquivo original foi convertido para texto bruto (plain text), mantendo a estrutura visual de indentação e quebras de página. Em seguida, foi aplicado um pré-processamento para limpeza de caracteres de controle (como Form Feed/Quebra de Página) e normalização de codificação de caracteres (UTF-8) para corrigir falhas de acentuação presentes na fonte (ex: correção de "8 -reas" para "8 Áreas").

- **Análise do Plano de Classificação (Estrutura Hierárquica)**

O script foi programado para identificar e extrair primeiramente o "Plano Geral de Classificação", reconhecendo padrões numéricos de identificação (IDs de categorias como 1, 1.1,

1.1.1, etc.). Foi implementada uma rotina de sanitização de nomes para remover caracteres que conflitavam com a sintaxe de importação do Tematres (substituição de "Termo: Qualificador" por "Termo - Qualificador"), garantindo que termos compostos (ex: "6.1.2 Publicações científicas: periódicos") fossem preservados integralmente.

- **Processamento da Ordem Alfabética e Linkagem**

Na etapa de varredura dos termos alfabéticos, o algoritmo realizou o cruzamento de dados com a estrutura hierárquica extraída anteriormente.

- **Detecção de Relações**

O script converteu automaticamente as notações originais (TG, TE, TR, CAT) para o padrão aceito pelo Tematres (BT, NT, RT, BT adicional). As notas de escopo, etiquetadas no TBCI como “ESP:” foram convertidas para a notação do TemaTres, “NA”; as versões do termo em idiomas inglês e espanhol, identificadas por ING: e ESP: foram configuradas como tipos de notas com as mesmas etiquetas no TemaTres.

- **Resolução de Orfãos**

Termos que apareciam isolados devido a quebras de página ou formatação inconsistente no original foram recuperados através de lógica de detecção de contexto (linhas em branco implícitas).

- **Normalização de Referências**

As referências a Termos ->Genéricos (Broader Terms) que utilizavam códigos numéricos ou grafias corrompidas foram normalizadas para seus nomes canônicos completos (ex: convertendo referências a "1.4" diretamente para "1.4 Métodos de Pesquisa e Análise"), garantindo a integridade da árvore taxonômica.

- **Tratamento automatizado**

Como resultado do tratamento automatizado, obteve-se um arquivo estruturado ("tesauro_final.txt") contendo a totalidade dos termos do TBCI, pronto para importação no Tematres. Foi usada a opção do TemaTres de importar usando texto etiquetado

Os principais resultados técnicos alcançados foram:

- Integridade Hierárquica: A estrutura de poli hierarquia foi preservada em até 4 ou mais níveis de profundidade (Ex: 1 -> 1.4 -> 1.4.1), resolvendo problemas anteriores onde subcategorias apareciam erroneamente como termos de topo.
- Eliminação de Duplicidades e Ruídos: O algoritmo de normalização eliminou variantes duplicadas criadas por erros de codificação (ex: unificação de "8 -reas" e "8 Áreas" em um único termo válido), resultando em um índice limpo.
- Recuperação de Termos Deslocados: Termos complexos ou situados em quebras de página, como "acordos de utilização" e "Publicações científicas - periódicos", foram corretamente posicionados em suas respectivas classes, eliminando a ocorrência de "termos órfãos" ou desconectados da árvore principal.

O tesauro final importado apresenta navegação fluida e hierarquia visualmente coerente, validada por testes de integridade nos nós críticos da taxonomia (Classes 1 a 8).

Além destes resultados, uma vez importados para o TemaTres, tanto a TAC quanto o TBCI podem ser exportados para o formato padrão SKOS, que é o objetivo deste experimento.

A exportação (TAC) para o formato SKOS (*Simple Knowledge Organization System*) representa um avanço significativo para a interoperabilidade de dados na pesquisa científica. Ao contrário do formato PDF tradicional, que funciona como um "depósito estático" de informações de difícil extração, o SKOS permite que a estrutura taxonômica seja lida e processada automaticamente por máquinas, integrando-se ao ecossistema da Web Semântica.

Essa disponibilidade em formato aberto e estruturado facilita a recuperação de informações e o cruzamento de dados entre diferentes sistemas de gestão acadêmica e repositórios digitais. Com o documento padronizado em SKOS, pesquisadores e desenvolvedores podem utilizar vocabulários controlados de forma dinâmica, garantindo que as áreas do conhecimento sejam vinculadas de maneira precisa e fiquem permanentemente acessíveis na rede como Dados Abertos Conectados.

4.3 Verificação da importação

A etapa de validação da integridade dos dados importados foi conduzida através da execução de scripts específicos de auditoria. Foram utilizados os comandos contidos no arquivo `verificar_tematres.sql` diretamente no banco de dados do sistema Tematres. Esses comandos permitiram uma análise detalhada da estrutura hierárquica, confirmando não apenas a contagem total de registros na tabela `lc_tema`, mas também a correta definição das Grandes Áreas como termos raiz (sem ascendência hierárquica) na tabela `lc_tabla_rel`.

Adicionalmente, foram realizadas verificações de consistência para identificar possíveis termos órfãos — registros que não possuíam vínculo hierárquico adequado e não eram classificados como raiz. A integridade estrutural foi duplamente checada cruzando as informações da tabela principal de temas com a tabela de índices (`lc_indice`), assegurando que todos os termos importados estivessem devidamente indexados para busca e navegação no sistema.

Ao final do processo de auditoria, confirmou-se que o total de 1335 termos foi importado e validado com sucesso. A estrutura hierárquica reproduziu fielmente a organização da Tabela de Áreas do Conhecimento da CAPES, com as 9 Grandes Áreas corretamente posicionadas no topo da hierarquia e seus respectivos desdobramentos em Áreas, Subáreas e Especialidades refletidos sem perdas ou inconsistências. Os dados também foram submetidos a um teste manual em uma amostragem de 10% do total, processo que consistiu na captura do texto original e na conferência direta via busca dentro do tesauro no TemaTres para garantir a consistência da conversão.

5 Conclusões

A conversão do formato original em PDF para o formato textual capaz de ser importado pelo TemaTres também trouxe ensinamentos importantes. A literatura sobre Big Data e Smart Data (IAFRATE, 2015) frequentemente enfatiza o valor extraído dos dados, mas por vezes subestima o esforço de engenharia necessário para tornar esses dados processáveis ("Machine Actionable"). O presente estudo de caso demonstrou que a conversão de um tesauro legível por humanos (PDF/Texto) para um formato semântico (SKOS) revela um "custo oculto" significativo

de pré-processamento. Inconsistências triviais para a leitura humana — como a codificação incorreta de caracteres ("8 -reas" em vez de "8 Áreas") ou variações sutis na notação de hierarquia ("Termo: Qualificador" vs "Termo - Qualificador") — constituem barreiras absolutas para a interoperabilidade computacional.

Uma descoberta acidental, porém, relevante deste projeto, foi o papel da migração automatizada como instrumento de auditoria da qualidade da informação. Ao submeter o TBCI a regras rígidas de validação hierárquica exigidas pelo software Tematres, foram identificados erros estruturais que permaneceram latentes na versão documental por anos. O caso do termo "acordos de utilização", que existia visualmente no documento original, mas estava logicamente desconectado da árvore hierárquica devido a uma quebra de página não tratada, exemplifica este fenômeno. A rigidez da lógica computacional (onde cada termo deve ter um pai e um identificador único) atuou como um filtro de qualidade (Quality Assurance), forçando a explicitação de relacionamentos tácitos.

Assim, argumenta-se que projetos de digitalização e migração para a Web Semântica cumprem uma dupla função: disseminação (acesso) e validação retrospectiva (integridade) do conhecimento organizacional.

Por fim, a opção pelo desenvolvimento de scripts de extração em linguagem Python (em detrimento da digitação manual dos termos) alinha este trabalho aos princípios da Ciência Aberta, especificamente no que tange à reprodutibilidade. A codificação das regras de tratamento em software documenta inequivocamente as decisões tomadas durante a curadoria: como foram tratados os caracteres especiais, quais critérios definiram a hierarquia e como foram normalizadas as variantes. Futuros pesquisadores que desejem atualizar ou auditar este tesauro não dependerão de manuais subjetivos, mas poderão examinar, executar e aprimorar o próprio código-fonte utilizado. Esta abordagem promove a transparência metodológica e garante que o processo de estruturação do conhecimento seja auditável e replicável, superando o modelo de "caixa preta" que frequentemente caracteriza migrações de dados proprietários.

Os tesauros que foram examinados para a realização deste experimento, o THESAURUS para acervos museológicos e a sua versão para a Web, o Tesauro do Folclore e Cultura Popular

THESAURUS de acervos científicos e o Tesauro de Cultura Material dos Índios do Brasil, mostram que a conversão do formato PDF para o formato TXT que servirá de entrada para o TemaTres é a parte crítica de um projeto como este. Os textos em PDF estão em layouts diversificados, o que indica que um tratamento específico para cada caso deve ser necessário. No experimento realizado foi necessário programação em Python para realizar esta tarefa.

6 Considerações finais

Soluções técnicas para transformar importantes tesauros no cenário brasileiro de Cultura, Memória e Patrimônio existem, como foi mostrado. Financiamentos públicos, como a recente Chamada Pública MCTI/FINEP/FNDCT RECUPERAÇÃO E PRESERVAÇÃO DE ACERVOS (Finep, 2024), têm como foco as instituições individualmente, e não políticas ou infraestruturas públicas e compartilhadas. No entanto, o desenvolvimento das TIs torna cada vez mais difícil, complexo e dispendioso que uma instituição individualmente consiga ter os recursos, tecnologia e expertise para realizar a curadoria de acervos digitais conforme descrito neste trabalho.

As instituições de Ciência e Cultura brasileiras estão bastante defasadas no desenvolvimento e uso de vocabulários. Não dispõem hoje de vocabulários padronizados compatíveis com o estado da arte no mundo, isto é, acessíveis via Web, com identificadores persistentes (Sayão, 2007) para seus conceitos, em formatos legíveis por máquinas (Marcondes, 2018), os LOV – Linked Open Vocabularies (Zeng, 2019), compatíveis com a proposta da Web Semântica e Dados Abertos Interligados. Os mais importantes vocabulários no cenário mundial de Cultura, Memória e Patrimônio vêm, desde há algum tempo, se alinhando com estas tecnologias. Exemplos significativos são os vocabulários desenvolvidos pela Fundação Paul Getty **(8)**, o ATT - Art and Architecture Thesaurus -, a CONA – Cultural Objects Name Authority -, a ULAN – Union List of Artists Names – e o TGN – Thesaurus of Geographic Names e os Linked Data Vocabularies for Cultural Heritage **(9)**, vocabulários como o Iconclass **(10)**, um vocabulário de tipos de imagens e iconografia, o UNESCO Thesaurus **(11)**, o VIAF – Virtual International Authority File **(12)**, entre outros.

Vocabulários importantes para a Ciência e Cultura brasileiras carecem de uma infraestrutura de TI para serem desenvolvidos, mantidos, consultados, compartilhados e poderem ser extraídos, exportados e reusados. Se estivessem disponíveis em formatos interoperáveis como SKOS, poderiam ser reusados, ser objeto de novos desenvolvimentos ou extensões para aplicações específicas. No contexto da Web, das Humanidades Digitais e do Big Data, não existe uma política pública brasileira para o desenvolvimento de instrumentos semânticos para dados digitais em Ciência e Cultura como são os vocabulários.

São necessárias soluções em termos de políticas públicas. Vocabulários são a chave para que estas instituições possam dispor dos meios de atribuir semântica a seus dados e processá-los usando as tecnologias de informação e Big Data. Estas devem contemplar identificação e criação vocabulários estruturantes no cenário brasileiro, criação de uma infraestrutura tecnológica pública que contemple repositórios, atribuição de identificadores persistentes para os objetos dos acervos digitais, questões de copyright e licenças de uso, interoperabilidade, formatos comuns e portais.

O experimento mostrou que a recuperação de vários vocabulários importantes para a Ciência e Cultura no Brasil é viável tecnologicamente. Sua disponibilização na Web em um formato portável como o SKOS, no escopo de uma política pública, pode viabilizar seu reuso em larga escala e seu desenvolvimento. Desta maneira, os diversos vocabulários já desenvolvidos e os ainda por desenvolver se tornarão “vivos” para serem usados e reusados no acesso e processamento dos dados brasileiros em Ciência e Cultura.

Notas

(1) Ver site em https://dados.gov.br/home

(2) Citação tirada do sítio http://site.mast.br/hotsite_museologia/thesaurus.html em 24/04/2025; o link para acessar o thesaurus está quebrado.

(3) A iniciativa 5STAR Open Data (2025) chama a atenção para as desvantagens dos dados não estruturados, "torne seus recursos disponíveis como dados estruturados (ex. excel no lugar de imagem escaneada)"

(4) Ver site em https://acervo.bn.gov.br/sophia_web/busca/autoridades; consulta realizada em 21/04/2025 com o termo "Brasil – História" recuperou 283 Termos Tópicos.

(5) Por exemplo, Vila Rica é o antigo e original nome da cidade de Ouro Preto, antiga capital do estado de Minas Gerais, conjunto histórico e arquitetônico tombado pela UNESCO. Pois o termo "Vila Rica" não aparece no TGN – Thesaurus of Geographic Names – da Fundação Paul Getty, consultado em 17 set. 2022.

(6) Ver em http://www.infolegis.com.br/lista-tesauros.html, consultado em 25 ago. 2022.

(7) Ver os resultados de uma pesquisa no Google com a seguinte expressão de busca: "cabeçalhos de assunto biblioteca universidade"

(8) Ver site com vocabulários em https://www.getty.edu/research/tools/vocabularies/lod/index.html

(9) Ver site com vocabulários em https://www.heritagedata.org/

(10) Ver site com vocabulários em https://iconclass.org/

(11) Ver site com vocabulários em http://vocabularies.unesco.org/browser/thesaurus/

(12) Ver site com vocabulários em https://viaf.org/en

Agradecimentos

Este trabalho foi realizado com o apoio das agências de fomento brasileiras CAPES - Código de Financiamento 001, e CNPq, processo nº 305253/2017-4.

Referências

5STAR Open Data. *5STAR Open Data*, 2025, https://5stardata.info/pt-BR/.

Baca, Murtha. "Prefácio." *Vocabulários Controlados: Terminologia para Arte, Arquitetura e Outras Obras Culturais*, by Patricia Harpring, Secretaria de Estado de Cultura / Pinacoteca de São Paulo, 2016, pp. 19–21.

Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." *Scientific American*, vol. 284, no. 5, May 2001, pp. 34–43, http://www.scian.com/2001/0501issue/0501berners-lee.html.

Dextre Clarke, Stella G., and Marcia Lei Zeng. “From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modelling.” *Information Standards Quarterly*, vol. 24, no. 1, 2012, https://eprints.rclis.org/16818/1/SP_clarke_zeng_isqv24no1.pdf.

Santos, José Carlos Francisco dos, Brígida Maria Nogueira Cervantes, and Mariângela Spotti Lopes Fujita. “Tesauro Eletrônico: Importação no TemaTres e Disponibilização na Web.” *XIX Encontro Nacional de Pesquisa em Ciência da Informação (XIX ENANCIB)*, 2018a.

Santos, José Carlos Francisco dos, Walter Moreira, and Mariângela Spotti Lopes Fujita. “Tesauro Unesp: Integração do Registro de Autoridade para o TemaTres.” *XIX Encontro Nacional de Pesquisa em Ciência da Informação*, 2018b.

Eletrobras. Tesauro do Setor de Energia Elétrica. Eletrobras, 1992.

Ferrez, Helena Dodd, and Maria Helena Soutello Bianchini. *Thesaurus para Acervos Museológicos*. Fundação Nacional Pró-Memória, 1987.

Finep. *Chamada Pública MCTI/FINEP/FNDCT/Identidade Brasil – Recuperação e Preservação de Acervos 2024*. FINEP, 2024, https://finep.gov.br/images/chamadas-publicas/2024/11_07_2024_Edital_Acervos.pdf.

Granato, Marcus, et al. “Thesaurus de Acervos Científicos em Língua Portuguesa: Concepção e Resultados Preliminares.” *ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação*, Rio de Janeiro, 2010, Ibict, https://ridi.ibict.br/bitstream/123456789/288/1/RosaliEnancib2010.pdf.

Iafrate, Fernando. *From Big Data to Smart Data*. Londres: ISTE, 2015

Marcondes, Carlos H., and E. M. Souza. “Vocabulários e Acesso Integrado a Acervos Digitais em Memória e Cultura.” *Seminário Internacional de Políticas Culturais*, Rio de Janeiro, 2018, pp. 109–24, https://pt.slideshare.net/slideshow/vocabularios-e-descricao-de-objetos-museologicos-para-recuperacao-da-informacao-na-web/278450826.

Marcondes, Carlos H. “Vocabulários e Descrição de Objetos Museológicos para Recuperação de Informações na Web.” *SlideShare*, 2018, https://pt.slideshare.net/slideshow/vocabularios-e-descricao-de-objetos-museologicos-para-recuperacao-da-informacao-na-web/278450826.

Motta, Dilza Fonseca da. *Tesauro de Cultura Material dos Índios do Brasil*. Museu do Índio/FUNAI, 2006.

NISO. *NISO TR-06-2017: Issues in Vocabulary Management*. NISO, 2017, https://www.niso.org/publications/tr-06-2017-issues-vocabulary-management.

Oliveira, Nirlei Maria, Maria das Dores Rosa Alves, and Gilmar Vicente. “O Cabeçalho de Assunto da Rede Bibliodata/Calco: Uso e Recuperação na Base Acervus/Unicamp.” *Transinformação*, vol. 9, no. 1, 1997.

Sayão, Luís Fernando. “Interoperabilidade das Bibliotecas Digitais.” *Transinformação*, vol. 19, 2007, pp. 65–82.

Shahrokni, Ali, and Jan Söderberg. “Beyond Information Silos.” *BigMDE 2015*, 2015, p. 63.

“UNESCO Thesaurus.” *UNESCO Vocabularies*, 3 May 2022, http://vocabularies.unesco.org/browser/thesaurus/.

“VIAF: The Virtual International Authority File.” *VIAF*, 15 Apr. 2020, https://viaf.org/en.

Will, Leonard. “The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus.” *Bulletin of the American Society for Information Science and Technology*, vol. 38, no. 4, 2012, pp. 48–51, https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/bult.2012.1720380413.

Zeng, Marcia Lei. “Interoperability.” *Knowledge Organization*, vol. 46, no. 2, 2019, pp. 122–46

Submetido: 24/01/2026 Aceito: 19/02/2026