em questão
E-ISSN 1808-5245

Proposta de uma ferramenta para classificação arquivística com base em ontologias

Daniel Libonati Gomes

Mestre; Universidade Federal do Pará, Belém, PA, Brasil;
daniellibonati00@hotmail.com

Thiago Henrique Bragato Barros

Doutor; Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil;
bragato.barros@ufrgs.br

Renato Tarciso Barbosa de Sousa

Doutor; Universidade de Brasília, Brasília, DF, Brasil;
renasou@unb.br

Roberto Lopes dos Santos Junior

Doutor; Universidade Federal do Pará, Belém, PA, Brasil;
robertolopes@ufpa.br

Resumo: Buscou-se construir e demonstrar uma ferramenta desenvolvida a fim de reduzir o aspecto subjetivo inerente à classificação arquivística, tornando-a mais consistente. Tendo-se em conta que erros de classificação podem prejudicar a grande maioria das funções arquivísticas, especialmente a avaliação e a descrição, foi elaborado um *software* denominado *Ontological Classifier* (OntoClass). Esse *software*, por meio da criação de uma ontologia a partir do plano de classificação de uma entidade produtora de documentos, é capaz de determinar a classe à qual um documento pertence com base em termos autorizados dispostos em uma lista. A fundamentação teórica foi realizada por meio de uma pesquisa bibliográfica e o desenvolvimento da ferramenta foi feito com uso da linguagem de programação Python 3.7 e da linguagem de consulta SPARQL. A partir de testes realizados com uma ontologia simples criada especificamente para este trabalho, conclui-se que o OntoClass alcança seu objetivo, apesar de ainda serem necessários testes em situações reais e apesar de haver alguns requisitos a cumprir para obter resultados positivos de sua utilização.

Palavras-chave: Arquivologia. Classificação. Ontologia. Linguagem de consulta. SPARQL.

1 Introdução

Na prática arquivística, a observância aos princípios arquivísticos, como o da proveniência, o da organicidade e o da unicidade, é fundamental para a organização de um acervo documental, para o acesso a ele e para o seu uso. Seguindo tais princípios, têm-se documentos pensados com base nas relações orgânicas que estabelecem uns com os outros por meio da acumulação documental no decorrer da execução das atividades de uma entidade, seja uma instituição ou uma pessoa. Porém, para que haja a organização de acordo com os princípios, é necessária a execução apropriada das funções arquivísticas, dentre as quais se destaca a classificação, tida como matricial por Sousa (2003).

A classificação é uma função matricial do trabalho arquivístico por fundamentar a realização de outras atividades de gestão de documentos, como a avaliação e a descrição. Além disso, a classificação assegura que os documentos sejam organizados de acordo com a proveniência e organicidade, pois agrupa, em um mesmo conjunto, documentos relacionados. Essa função garante ainda uma forma de recuperação da informação por meio de seu produto, comumente chamado de plano de classificação, que representa os conjuntos documentais e atribui a eles códigos que permitam sua identificação.

No entanto, justamente por conta do caráter fundamental da classificação, é importante que não haja erros na realização dessa função. Como Sousa (2014) deixa claro, é muito comum que os responsáveis por classificar os documentos não sejam arquivistas e não tenham o conhecimento adequado para realizar essa atividade, o que pode levar a uma má compreensão do plano de classificação e, consequentemente, resultar em erros de classificação. Isso inclusive leva esse autor a defender que uma das funções do arquivista é construir instrumentos inteligíveis para os usuários. Sem isso, corre-se o risco de serem agrupados determinados documentos que deveriam estar separados ou de separar documentos que deveriam estar juntos. Assim, fica evidente que a classificação é, de certa forma, uma atividade subjetiva, que depende da interpretação pessoal de um documento.

Diante desse fato, tem-se o objetivo do presente trabalho: propor uma ferramenta que facilite a realização da classificação, tendo como base características dos documentos. Para tanto, optou-se pelo uso das ontologias,

Sistemas de Organização do Conhecimento (SOC) (VITAL; CAFÉ, 2011; CARLAN; MEDEIROS, 2011) que organizam conceitos relativos a domínios de conhecimento em classes. Essa opção se deu pelo fato de que, como foi demonstrado por Barros e Gomes (2018), as ontologias podem ser utilizadas para auxiliar no desenvolvimento de planos de classificação mais consistentes, sendo construídas anteriormente a estes e possibilitando, com isso, a visualização mais clara dos elementos que devem compô-lo. No presente trabalho, objetivou-se, portanto, levar essa ideia a outro nível a partir da elaboração de uma ferramenta que, por meio de ontologias, auxiliasse na atividade de classificação.

As ontologias são ferramentas poderosas, utilizadas para diversos fins, como a interoperabilidade entre sistemas diferentes ou a simples representação de domínios de conhecimento. Vale frisar que tais sistemas de organização podem ser desenvolvidos de maneira mais simples, com um modelo visual que permite a ligação entre conceitos de um domínio, como também podem ser desenvolvidos em linguagens específicas de programação, sendo a linguagem *Ontology Web Language* em sua segunda versão, chamada OWL2, recomendada atualmente pelo *World Wide Web Consortium* (W3C).

Outra característica das ontologias, especificamente daquelas desenvolvidas para sistemas computacionais, é a capacidade de serem "consultadas", isto é, é possível fazer questionamentos a uma ontologia, dependendo da forma como é organizada. Por exemplo, em uma ontologia de animais, seria possível perguntar quais animais são mamíferos, de modo que o resultado seria o nome de todos os mamíferos inseridos no sistema. Isso é possível por meio das chamadas linguagens de consulta (*query languages*, em inglês), dentre as quais se podem destacar as linguagens SPARQL *Protocol and RDF Query Language* (SPARQL) e *Semantic Query-Enhanced Web Rule Language* (SQWRL). Neste trabalho será dado foco à linguagem SPARQL.

Portanto, tendo em conta a possibilidade de consultas às ontologias e a representação de um plano de classificação nesses sistemas, pensou-se na elaboração de um *software* que permitisse a realização da classificação tendo como base termos pré-selecionados relativos a características dos documentos

produzidos por uma instituição ou pessoa. O programa deveria possibilitar que, após a análise do documento, o usuário pudesse selecionar determinados termos (já presentes em uma lista preparada juntamente ao plano de classificação) e, inserindo-os em um campo específico, tivesse como retorno a classe específica à qual o documento pertence. Diante disso, desenvolveu-se o *software Ontological Classifier* (OntoClass).

Para a elaboração do OntoClass, inicialmente foi realizada uma pesquisa bibliográfica nas bases de dados Google Acadêmico e Portal de Periódicos CAPES voltada a obter informações gerais acerca da classificação arquivística, de sua importância e de sua execução, bem como informações sobre ontologias e, principalmente, linguagens de consulta. Os estudos sobre classificação foram utilizados como base para pensar a estratégia de funcionamento do *software*, enquanto os estudos sobre ontologias permitiram tornar essa estratégia executável.

Nas seções seguintes, foi inicialmente dado foco à classificação arquivística, a seus métodos e a requisitos a que um plano de classificação deve atender; em seguida, foi feita uma exposição acerca das ontologias e linguagens de consulta; por fim, foi demonstrado o OntoClass, assim como suas características e as exigências que devem ser atendidas para garantir sua devida utilização.

2 A classificação arquivística

O arquivista desempenha um conjunto de funções denominadas funções arquivísticas, que são: criação/produção, aquisição, conservação/preservação, avaliação, classificação, descrição e indexação e difusão/acesso (ROUSSEAU; COUTURE, 1998). Dentre estas, porém, uma se destaca por seu caráter matricial, possibilitando e facilitando a realização de outras: a classificação (SOUSA, 2014).

A classificação pode ser entendida como a divisão de um todo em partes, chamadas classes, com base nas características semelhantes e diferentes de cada uma dessas partes. Dessa forma, a classificação arquivística pode ser conceituada como o processo de organizar, de maneira hierárquica, um dado

acervo em classes e subclasses, conforme as características dos documentos, envolvendo tanto seus elementos formais, como o suporte, quanto os relativos ao conteúdo.

No entanto, a classificação arquivística tem como diferencial de outros tipos de classificação (como a bibliográfica) o seu objeto, o documento de arquivo. Os arquivos têm algumas características fundamentais, como o fato de deverem sempre ser compreendidos em seu todo, sendo que esse todo é maior que a soma de suas partes. Assim, a classificação arquivística procura organizar um objeto que não pode ser compreendido completamente a partir de suas unidades, mas, sim, em conjuntos maiores. Na prática, isso se reflete no fato de os documentos só terem realmente sentido quando relacionados a outros, seja no âmbito dos arquivos correntes e intermediários, seja no dos arquivos permanentes (fase em que a classificação, na literatura, geralmente é chamada de arranjo, mas que segue as mesmas premissas, tratando-se, portanto, de uma convenção, sem nenhum reflexo teórico-metodológico). No entanto, é necessário sublinhar que, em perspectivas mais recentes, a compreensão da individualidade dos documentos também tem sido vista como possível, especialmente se for levado em consideração o entendimento de proveniência trazido pela Arquivologia Pós-moderna (COOK, 2012).

Diante disso, é importante deixar claro que a classificação atuará sempre tendo em conta os princípios arquivísticos, que, segundo Bellotto (2002), são: princípio da proveniência (ou respeito aos fundos), princípio da organicidade, princípio da indivisibilidade, princípio da cumulatividade e princípio da unicidade. Dentre estes, destacam-se a proveniência, a organicidade e a unicidade, tendo em conta que a indivisibilidade e a cumulatividade são consequências dos primeiros.

O princípio da proveniência se refere ao fato de que todo documento de arquivo tem sua identidade fixada à de seu produtor, ou seja, toda organização deve ser feita levando-se em conta essa relação entre documento e produtor, além de que a ordem original em que os documentos são produzidos também é importante por refletir a realização das atividades que os originaram. Partindo daí, o princípio da organicidade diz respeito ao fato de que o acervo de uma

dada instituição reflete diretamente as ações administrativas realizadas por ela, isto é, por meio do arquivo é possível compreender diversos aspectos da própria instituição. Por fim, o princípio da unicidade – que se aplica mais a documentos convencionais do que aos digitais – releva um fato acerca dos documentos tomados individualmente: todo documento de arquivo, não importando seu gênero, tipo ou suporte, é único, tendo em vista seu contexto de produção.

Partindo desses princípios, é evidente aquilo que o arquivista deverá considerar para a realização da classificação. Primeiramente, esta deve ser feita tendo em conta a proveniência dos documentos, de maneira a não misturar os documentos de uma origem com outros de outra. Sousa (2007) afirma que a proveniência é uma característica essencial dos documentos de arquivo, visto que estes só têm a forma e o valor que têm por conta das atividades realizadas pelo seu produtor, sendo, portanto, a proveniência uma marca da identidade dos documentos de arquivo. Além disso, esse princípio também orienta a manter a ordem original em que os documentos foram produzidos, o que é importante para as divisões do fundo arquivístico propostas na classificação.

Tendo em conta a proveniência, são perceptíveis os outros princípios arquivísticos no processo de classificação. Classificar respeitando a proveniência fará com que a classificação reflita a cumulatividade natural dos documentos e, consequentemente, a organicidade do acervo. Além disso, dispor os documentos em classes também torna evidente sua indivisibilidade, ou seja, demonstra que os documentos desse arquivo devem seguir a classificação estabelecida, com o risco de, caso isso não aconteça, perderem-se materiais importantes para o uso corrente ou para pesquisas, no caso dos arquivos permanentes, o que também destaca a unicidade dos documentos de arquivo.

A classificação pode ser operacionalizada, dando origem a um instrumento denominado plano de classificação (ou quadro de arranjo, no caso dos arquivos permanentes), que evidenciam quais são as classes em que o arquivo se divide. O plano de classificação é fundamental para a realização de diversas atividades nos arquivos, envolvendo tanto o controle do acervo quanto a recuperação das informações. Mais especificamente, é a partir do plano de classificação que outras funções arquivísticas, como a avaliação e a descrição,

são realizadas. A primeira se caracteriza pela análise dos documentos para estabelecer os prazos de guarda e a destinação, tendo em conta os valores que lhes são atribuídos; a segunda busca representar o acervo como um todo, destacando suas informações tanto institucionais como de conteúdo dos documentos a fim de permitir a produção de instrumentos de pesquisa (guia, catálogo, inventário etc.).

No caso da avaliação, Sousa (2014) afirma que as avaliações não são feitas de documento a documento, de maneira que o plano de classificação ajuda a situar a estrutura hierárquica dos conjuntos documentais, informação essa que é muito importante para o processo de avaliação. Já no caso da descrição, a classificação auxilia na obtenção de informações essenciais para uma descrição eficaz e, consequentemente, a elaboração de bons instrumentos de pesquisa, além de que a atividade de descrição é feita sempre levando-se em consideração os níveis em que o acervo se divide.

Além de tudo isso, é importante ainda ressaltar que a classificação é fundamentada por um estudo geral da entidade produtora dos documentos, buscando compreender desde a missão até as características dos documentos produzidos. Por meio desse estudo, é possível compreender que entidade produz e acumula os documentos, em que contexto ela está inserida, por que produz determinados tipos documentais e como faz isso. Assim, obtém-se o material necessário para construir um plano de classificação eficiente que realmente represente as partes em que se divide a entidade.

Feitas as pesquisas sobre a entidade produtora, pode-se pensar no método de classificação a ser adotado, isto é, o fundamento no qual os documentos serão organizados na classificação. Conforme Schellenberg (2006), a classificação tradicionalmente pode ser funcional, estrutural ou por assunto. Em geral, as classificações são feitas com base nas funções da organização, como classes maiores, e suas atividades, como subclasses. Segundo Gonçalves (1998), as classificações funcionais tendem a atender melhor às exigências da classificação arquivística – apesar de esse método apresentar algumas dificuldades de aplicação, como já explicado anteriormente, de maneira que se pode argumentar que ele tende a ser mais eficiente para os arquivistas do que

para os usuários. Já as classificações estruturais, que têm por base a estrutura organizacional, são mais indicadas nos casos de instituições com estruturas mais estáveis e funções bem definidas. Por fim, a classificação por assunto só deve ser elaborada em casos específicos, para documentos com função de prover referências ou informações relativas a assuntos muito específicos.

Outro aspecto ainda a ser considerado é a classificação dos documentos no âmbito digital, realizada na gestão de documentos digitais por meio de um Sistema Informatizado de Gestão Arquivística de Documentos (SIGAD), que segue as diretrizes dadas pelo Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos (e-ARQ Brasil). Conforme Schäfer e Lima (2012), a classificação de documentos em SIGAD tem a vantagem de não ser opcional, fazendo com que o "descaso" com a classificação dos documentos seja minimizado. Além disso, vale salientar que a classificação no ambiente digital ainda obedece a um plano de classificação, visto que um SIGAD possui funcionalidades que abrangem toda a gestão documental, o que inclui a avaliação. Dessa forma, um documento precisa, mesmo no meio digital, ser classificado corretamente para que sua gestão seja feita de forma devida. Ainda segundo esses autores (SCHÄFER; LIMA, 2012):

> [...] as implicações da classificação equivocada são as mesmas que as no meio físico. Por mais que o sistema seja evoluído tecnologicamente, a tarefa de classificar os documentos permanece como atividade inerente ao indivíduo. Reitera-se, portanto, que a tecnologia não é capaz de dirimir a importância da compreensão por parte dos envolvidos, não só para a classificação, mas para toda a gestão documental de uma Instituição. (SCHÄFER; LIMA, 2012, p. 149)

Portanto, mesmo com a elaboração eficaz de um plano de classificação, seja ele aplicado no âmbito convencional ou no digital, ainda é necessário que a atividade de classificar seja realizada de forma correta e consistente. Documentos classificados erroneamente poderão ter sua avaliação e descrição afetadas, além de serem colocados em conjuntos documentais aos quais não pertencem. Esse erro implica a possibilidade de um documento receber uma destinação e prazos de guarda errôneos ou não estar bem contextualizado para sua descrição.

Tendo isso em conta, vale lembrar que, conforme Sousa (2014) há situações em que os responsáveis pela classificação de documentos de arquivo, além de não serem arquivistas, também têm dificuldade em compreender o plano de classificação utilizado, especialmente se este tomar como princípio as funções e atividades realizadas pela instituição – a já explicada classificação funcional. Deve-se ter em mente que, em geral, as pessoas não pensam nas funções realizadas por um documento, mas, sim, nos assuntos sobre os quais trata ou em sua forma.

Portanto, diante da grande importância da classificação e das consequências trazidas pelos erros em sua execução, é necessário que sejam pensadas soluções que reduzam as chances de tais erros ocorrerem, seja por descuido do responsável por essa atividade, seja pela subjetividade que a permeia na análise de um plano de classificação e no contato com um documento a ser classificado. Assim, na seção seguinte, trata-se das ontologias, sistemas de organização por meio dos quais propomos uma solução para o problema aqui descrito.

3 Ontologias e a organização do conhecimento

As ontologias, assim como outros SOCs, como os tesauros, índices e planos de classificação, são estudadas pela Ciência da Informação como ferramentas para representar a informação e possibilitar sua recuperação. No entanto, as ontologias sobressaem atualmente por serem peça fundamental na chamada Web Semântica, projeto que visa a estabelecer o compartilhamento e reuso de dados diversos em aplicações variadas. Por meio desses sistemas de organização, podem-se criar taxonomias de conceitos relativos a um domínio de conhecimento específico, possibilitando que um sistema computacional "entenda" esse domínio, levando-se em conta que, para tal sistema, o que existe é aquilo que pode ser representado (GRUBER, 1993).

Há diversos conceitos para ontologia e não é fácil estabelecer completamente um, considerando que existem muitas nuances nessa discussão. Neste trabalho, tomaremos o conceito proposto por Guarino (1997): "Uma ontologia é uma descrição explícita e parcial dos modelos pretendidos de uma

linguagem lógica" (GUARINO, 1997, p. 298, tradução livre). No entanto, tal conceito é ainda bastante complexo e envolve muitas explicações, que devem ser fornecidas.

Para compreender o conceito de Guarino (1997), é necessário entender que uma ontologia descreve uma conceptualização, que pode ser compreendida como sendo o conjunto de conceitos e suas definições que estabelecem relações dentro de um campo de interesse qualquer. Porém, as conceptualizações, em geral, são implícitas, existindo de forma diferente na mente das pessoas. Assim, uma ontologia deve ser feita de modo contrário: a conceptualização nela representada deve ser explícita e clara. Além disso, o conceito aponta que essa conceptualização também deve ser parcial, ou seja, uma ontologia nunca vai representar um domínio de conhecimento em sua totalidade, mas apenas aqueles conceitos que são pretendidos.

Por fim, Guarino (1997) afirma que a ontologia atua por meio de uma linguagem lógica, ou seja, por meio de axiomas. Em suma, uma ontologia, por meio de axiomas, descreve e explicita algumas das relações existentes (apenas aquelas pretendidas) entre uma série de conceitos que formam um domínio de conhecimento (também chamado de universo discursivo). Na prática, isso ocorre por meio de uma estrutura taxonômica que conjuga três elementos principais: classes, propriedades e instâncias (ou indivíduos) (NOY; MCGUINNESS, 2001).

As classes de uma ontologia, em geral, representam os conceitos do domínio que se está buscando representar. As instâncias ou indivíduos são os elementos que se inserem no interior das classes. Por fim, as propriedades permitem que haja relações entre classes, entre indivíduos e entre classes e indivíduos. Conforme já foi dito na Introdução, uma ontologia pode ser desenvolvida em diversos formatos, desde modelos conceituais em gráficos até linguagens próprias de sistemas informatizados. A escolha de uma forma de representação varia de acordo com o objetivo que se busca alcançar com o desenvolvimento da ontologia. No caso deste trabalho, pretende-se a criação de ontologias formais, ou seja, aquelas representadas por linguagens próprias da Web Semântica, como a OWL2, já mencionada. Para tanto, pode-se desenvolver

o código diretamente ou usar editores como o *software* Protégé 5.2.0 (MUSEN, 2015), que foi escolhido para o desenvolvimento do exemplo que será utilizado a seguir.

3.1 Tornando planos de classificação em ontologias

Ontologias fornecem uma representação visual muito clara de um domínio de conhecimento, permitindo sua modificação e reuso a qualquer momento. Diante disso, conforme Barros e Gomes (2018), a construção de um plano de classificação pode ser facilitada com o uso de ontologias terminológicas, conferindo completude ao plano desenvolvido. Vale ainda frisar que as ontologias são voltadas ao reuso, além de serem flexíveis e de fácil modificação, de maneira que se houvesse qualquer necessidade de mudança nas funções ou na estrutura da entidade produtora dos documentos e se essa necessidade se refletisse no plano de classificação, a mudança poderia ser facilmente feita na ontologia. Seguindo essa linha de pensamento, elaboramos um plano de classificação simplificado, voltado apenas para exemplificação e teste do produto desenvolvido (descrito na seção 4).

O plano de classificação foi desenvolvido a partir de alguns documentos do fundo Objeto Voador Não Identificado (OVNI), presente no Sistema de Informações do Arquivo Nacional (SIAN). Além disso, as funções e atividades descritas são baseadas puramente em uma análise superficial dos documentos.

No entanto, isso não compromete o resultado final do que está sendo proposto neste trabalho, haja vista que, mesmo que o plano seja apenas exemplificativo, sua forma e suas características são equivalentes às de um plano real, além de que a ferramenta desenvolvida e explicada no final do trabalho independe da forma ou da completude do plano de classificação, sendo muito mais dependente da maneira como o plano é transformado em ontologia. Dito isso, o plano de classificação que será aqui utilizado como exemplo pode ser visualizado no quadro abaixo:

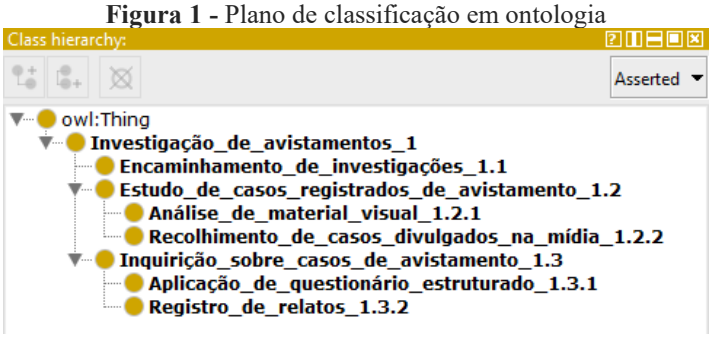
Quadro 1 - Plano de classificação

	OBJETO VOADOR NÃO IDENTIFICADO (OVNI)
1	**Investigação de avistamentos**
1.1	Encaminhamento de informações relativas a avistamentos
1.2	Estudos de casos registrados de avistamento
1.2.1	Análise de material visual
1.2.2	Recolhimento de casos divuldados na mídia
1.3	Inquirição sobre casos de avistamento
1.3.1	Aplicação de questionário estruturado

Fonte: Elaborado pelos autores (2018).

A passagem do plano em si para a ontologia não é nem um pouco complexa. O que seriam os conceitos, as classes na ontologia, se tornam as classes do plano de classificação. Caso fosse desejado, seria possível expressar relações diversas entre as classes do plano, porém esse não é o intento aqui. A ontologia que contém o plano pode ser visualizada abaixo:

Figura 1 - Plano de classificação em ontologia



Fonte: Elaborado pelos autores no Protégé 5.2.0 (2018).

A forma do plano e a divisão das classes são bastante claras, mas algumas características devem ser explicadas. Primeiramente, tem-se a classe **owl:Thing**, que é uma classe predeterminada na linguagem OWL e denota todos os elementos existentes no universo discursivo, ou seja, tudo que está na ontologia é caracterizado como uma **coisa**, *thing*. A segunda característica que chama a atenção é a escrita das classes, em que não há espaços (substituídos por pelo símbolo _, *underline*), e o código de classificação, que aparece ao final do nome da classe, dado que os elementos de uma ontologia não devem ter seu nome local iniciado com números.

Outro importante elemento dessa ontologia são as instâncias criadas para as classes, e é aí que se pode começar a compreender o mecanismo por trás da ferramenta apresentada na seção 4, o OntoClass. Para cada classe, foram selecionados termos que representem os documentos que podem ser nelas

classificados. Por exemplo, a classe **1.2.1 Análise de material visual** guarda documentos como fotografias e ilustrações, termos estes selecionados para essa classe. Assim, o quadro abaixo destaca as classes que recebem documentos e os termos a elas associados:

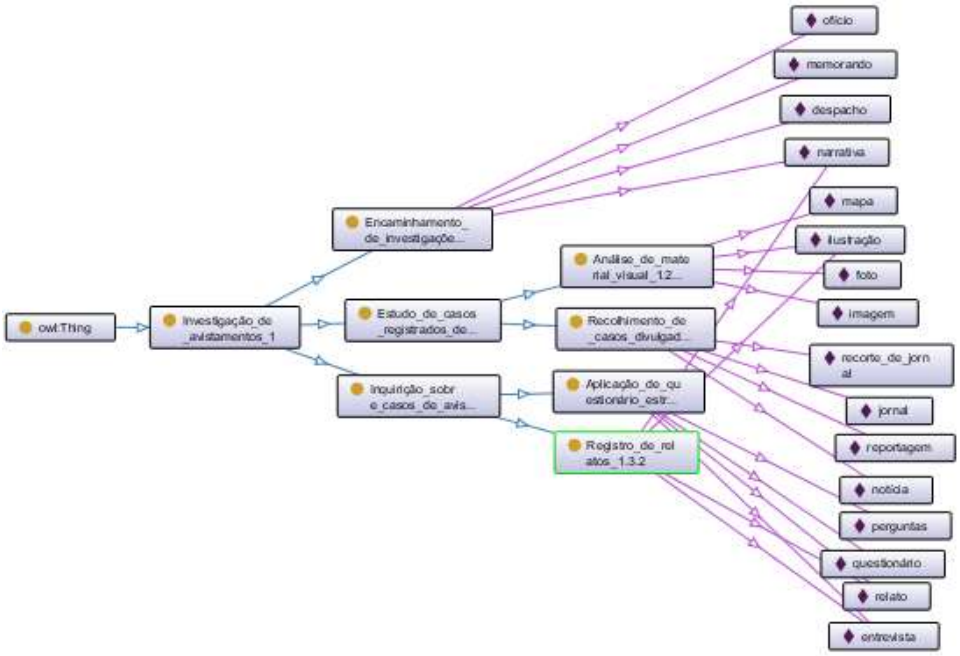
Quadro 2 - Classes do plano de classificação e termos relacionados

CLASSE	TERMOS
1.1 Encaminhamento de investigações	despacho, memorando, ofício, narrativa
1.2.1 Análise de material visual	foto, ilustração, imagem, mapa
1.2.2 Recolhimento de casos divulgados na mídia	jornal, notícia, recorte de jornal, reportagem
1.3.1 Aplicação de questionário estruturado	entrevista, perguntas, questionário, relato
1.3.2 Registro de relatos	entrevista, ilustração, narrativa, relato

Fonte: Elaborado pelos autores (2018).

A figura abaixo provê uma visão geral da ontologia, em que os elementos marcados em amarelo são as classes e os marcados em roxo são os termos (instâncias):

Figura 2 - Visão geral da ontologia baseada no plano de classificação



Fonte: Elaborado pelos autores no Protégé 5.2.0 (2018).

Como é possível perceber, os termos se relacionam a características dos documentos que podem ser adequados a cada classe e, para cada classe, foram escolhidos quatro termos a fim de evitar casos em que um termo leva a várias classes diferentes. Evidentemente, o que está proposto aqui não é uma regra. A

quantidade de termos escolhidos para cada classe vai variar em cada caso, e uma lista com os termos escolhidos deve ser elaborada em conjunto com o plano de classificação, quando da investigação da entidade produtora dos documentos.

A organização dessa lista pode, claro, variar, porém uma organização alfabética dos termos, divididos com base em assuntos ou características mais evidentes dos documentos, poderia render um bom resultado. Por exemplo, a lista de termos descritos para o caso aqui exemplificado poderia ser:

Quadro 3 - Lista de termos autorizados

Correspondência
despacho
memorando
ofício
Imagem
foto
ilustração
imagem
mapa
Narrativa
entrevista
jornal
narrativa
notícia
perguntas
questionário
recorte de jornal
relato
reportagem

Fonte: Elaborado pelos autores (2018).

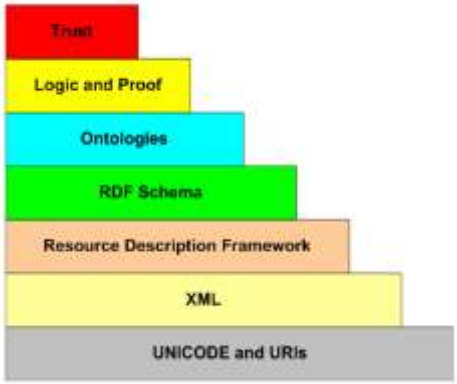
Dessa forma, o responsável por classificar os documentos, diante destes, com base nas características mais evidentes, poderia facilmente selecionar os termos correspondentes à classe do documento em questão. Porém, resta ainda a questão de como, por meio dos termos, chegar às classes. Isso pode ser realizado por meio das linguagens de consulta, sendo este, portanto, o tema da seção a seguir.

3.2 Consultando a ontologia desenvolvida

Conforme já dito anteriormente, as ontologias utilizam uma linguagem chamada OWL2 para expressar as relações semânticas entre entidades variadas. No entanto, para compreender como isso ocorre e quais são os elementos envolvidos nessas relações semânticas, é necessário entender um pouco do

funcionamento da Web Semântica. A figura abaixo demonstra a arquitetura geral da Web Semântica:

Figura 3 - Arquitetura da Web Semântica



Fonte: MATTHEWS (2005, p. 4).

Inicialmente, tem-se na base o Unicode e o URI (Uniform Resource Identifier), que foi expandido pelo IRI (Internationalized Resource Identifier). Ambos atuam como identificadores, de maneira que o Unicode atribui um número único para cada caractere utilizado, enquanto o IRI identifica o recurso propriamente dito, ou seja, é uma espécie de localizador. Por exemplo, um IRI de uma ontologia pode ser algo como **http://www.test.org /ontologies/2018/planoteste**. Cada entidade no interior da ontologia, seja uma classe, uma propriedade ou um indivíduo, tem como nome o IRI (que passa a ser visto como um prefixo) mais o nome local, de modo que, se esse IRI pertencesse à ontologia demonstrada na seção anterior, a classe **Investigação_de_avistamentos_1** teria o nome **http://www.test.org /ontologies/2018/planoteste#Investigação_de_avistamentos_1**.

A camada seguinte apresenta a XML (eXtensible Markup Language), uma linguagem que permite que toda a Web Semântica atue sob a mesma sintaxe. Essa mesma sintaxe é utilizada na linguagem OWL2.

Após a XML, acha-se o RDF (Resource Description Framework), que tem a função de representar de forma gráfica informações sobre recursos. O RDF trabalha com os trios sujeito-predicado-objeto, de maneira que permite que seja feita a relação entre dois recursos por meio de uma propriedade. O RDF

Schema expande essa ideia, provendo um vocabulário necessário para um modelo RDF, por meio de elementos como classes, subclasses, propriedades etc.

Por fim, na camada seguinte, é possível enfim ver as ontologias, que expandem o vocabulário disponibilizado em RDF e RDF Schema, possibilitando que relações semânticas sejam estabelecidas de fato. Por exemplo, uma ontologia consegue expressar que, se A é marido de B, então B é esposa de A, coisa que não é possível apenas com RDF. As camadas superiores às ontologias as levam mais além, estabelecendo relações lógicas que são passíveis de verificação por um processo de inferências e que podem ser consideradas confiáveis por meio de assinaturas digitais ou de outros meios fundamentados por agentes confiáveis.

Assim, diante desse quadro, é possível perceber que as ontologias, inclusive a que foi construída para este trabalho, são desenvolvidas em linguagem OWL2, que se apoia na sintaxe XML e que garante maiores funcionalidades para os recursos trazidos pelo RDF e RDF Schema. Dito isso, enfim é possível compreender o funcionamento das linguagens de consulta, especialmente da linguagem que trataremos aqui, o SPARQL.

Como o próprio nome SPARQL determina, a atuação dessa linguagem de consulta utiliza os trios RDF para fazer perguntas, que, no caso aqui estudado, são direcionadas a uma ontologia. Um modelo geral de consulta com SPARQL pode ser o seguinte:

Quadro 4 - Modelo de consulta com SPARQL

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX agenda:<http://agenda.com/test#>
SELECT ?nome ?email
WHERE { ?nome rdf:type agenda:Pessoa .
        ?nome agenda:tem_email ?email }
```

Fonte: Elaborado pelos autores (2018).

O código acima pedeque sejam mostrados os nomes das entidades da classe **Pessoa** e seus respectivos e-mails. O resultado dessa consulta seria uma tabela com duas colunas, uma com os nomes e outra com os e-mails.

Porém, alguns elementos devem ser destacados: *PREFIX*, *SELECT* e *WHERE*. O primeiro funciona como um substituto da IRI informada, ou seja, ao invés de escrever **http://www.w3.org/1999/02/22-rdf-syntax-ns#**, pode-se

apenas escrever **rdf:**, somado ao nome de algum elemento RDF, como **type**, que aparece no exemplo. O *SELECT* informa quais são os elementos que se está buscando, isto é, o que deve ser exibido no gráfico RDF. Enfim, o *WHERE* indica o que deve ser consultado, sendo nele presentes os trios RDF sujeito-predicado-objeto. O quadro abaixo explica o sentido de cada linha do código acima:

Quadro 5 - Explicação do Quadro 4

SPARQL	Explicação
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX agenda:<http://agenda.com/test#>	Informa os prefixos que substituirão as IRIs. No caso, "rdf:" pode ser usado para utilizar qualquer elemento da sintaxe RDF, enquanto que "agenda:" pode ser usado para se referir aos elementos da ontologia "http://agenda.com/test#".
SELECT ?nome ?email	Informa o que, dentro do *WHERE*, deve ser pesquisado, além dar nome às colunas que comporão a tabela RDF.
WHERE { ?nome rdf:type agenda:Pessoa . ?nome agenda:tem_email ?email }	Informa primeiro que "nome" é uma instância da classe "Pessoa", inserida na ontologia "agenda". Depois, afirma que os elementos "nome" estão ligados aos elementos "email" pela propriedade "tem_email", presente na ontologia "agenda".

Fonte: Elaborado pelos autores.

Essa mesma linguagem, portanto, pode ser usada para pesquisar as classes de um documento a partir dos termos informados, inseridos na ontologia criada a partir do plano de classificação como instâncias de suas classes. Assim, na seção a seguir, é apresentado o OntoClass e a mecânica utilizada para a consulta às ontologias.

4 A classificação arquivística por meio do Ontological Classifier

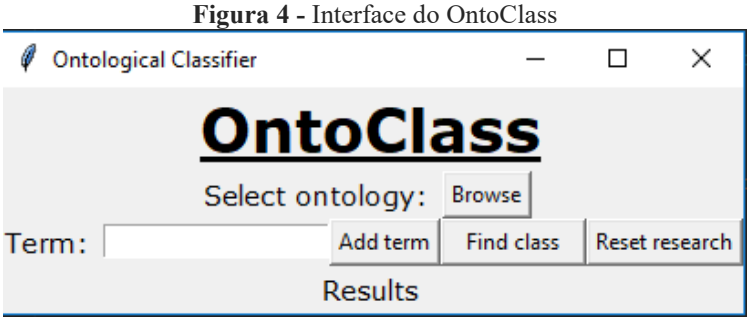
O desenvolvimento de um plano de classificação é uma tarefa complexa, que envolve uma análise cuidadosa de diversos aspectos da entidade produtora dos documentos. A missão, os objetivos, a estrutura, as funções, as atividades e os tipos documentais produzidos são todos elementos que devem estar bem claros para o arquivista. Assim, conforme foi dito anteriormente, para que a proposta deste trabalho tenha efeito positivo, é necessário que o plano de classificação

seja elaborado de forma eficiente, assim como a lista de termos característicos de cada classe.

O objetivo do *software* desenvolvido é produzir uma classificação consistente por meio dos termos pré-selecionados e organizados em uma lista. Evidentemente, a escolha dos termos é em si uma atividade subjetiva; porém, como tais termos são definidos e fundamentados em aspectos facilmente perceptíveis dos documentos, como tipologia ou espécie, a chance de erros tende a diminuir. Com isso, pode-se realizar uma classificação funcional com fundamento em aspectos mais superficiais dos documentos.

O OntoClass, em fase de protótipo, permite ao usuário escolher a ontologia que pretende utilizar, elencar termos de busca e, com isso, retorna a(s) classe(s) relacionada(s) ao(s) termo(s) escolhido(s). Evidentemente, a ontologia utilizada deve seguir os parâmetros apontados na seção 3.1 deste trabalho, de maneira que o plano de classificação deve ser desenvolvido antes dela. Assim, desde já se deixa claro que o OntoClass só poderá alcançar seu objetivo se o plano de classificação for desenvolvido de maneira adequada. A figura apresenta a interface do programa:

Figura 4 - Interface do OntoClass



Fonte: Elaborado pelos autores (2018).

A interface foi pensada para ser simples e intuitiva. É possível adicionar tantos termos quantos forem necessários e reiniciar a pesquisa. Os termos digitados e armazenados pelo botão **Add term** são registrados logo antes da área **Results**.

Para que o programa faça a pesquisa e identifique a ontologia, foi utilizado o módulo de programação em Python chamado Owlready2, que possibilita a construção e a leitura de ontologias. Como já mencionado

anteriormente, para a consulta a linguagem utilizada foi o SPARQL. O código que permite a leitura da ontologia selecionada pelo usuário e a consulta a ela pode ser visualizada abaixo:

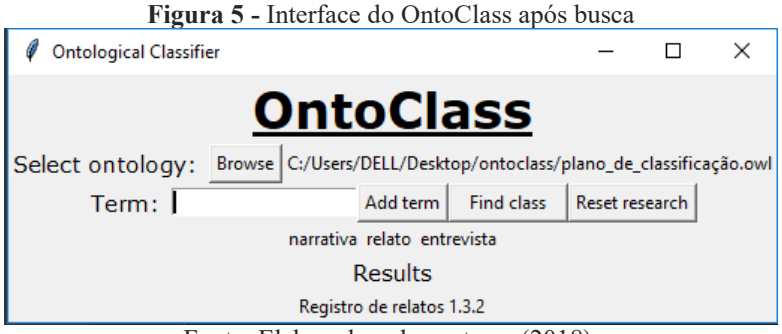
Quadro 6 - Reconhecimento da ontologia selecionada e consulta em SPARQL

```
1   self.ontology = World()
2   self.onto = self.ontology.get_ontology('file:///'+self.filename).load()
3   self.baseiri = self.onto.base_iri
4
5   self.graph = self.ontology.as_rdflib_graph()
6   self.query_parts = []
7   self.prefix = "PREFIX ont: <%s>" % self.baseiri
8   self.select_where = "SELECT ?class WHERE {"
9   self.query_parts.append(self.prefix)
10  self.query_parts.append(self.select_where)
11  for i in self.terms:
12      self.term = "ont:%s a ?class ." % str(i).replace(' ', '_')
13       self.query_parts.append(self.term)
14  self.closing = "}"
15  self.query_parts.append(self.closing)
16  self.request = "\n".join(self.query_parts)
17  self.results = list(self.graph.query(self.request))
```

Fonte: Elaborado pelos autores (2018).

Por meio do código acima, tem-se a leitura da ontologia (linha 2) e a consulta em SPARQL (linhas 5 a 17). Assim, após a seleção de termos de pesquisa, os resultados são exibidos na mesma janela da interface inicial, conforme a figura a seguir:

Figura 5 - Interface do OntoClass após busca



Fonte: Elaborado pelos autores (2018).

Como é possível perceber, o programa exibe o caminho da ontologia selecionada pelo usuário. Com três termos (**entrevista**, **relato** e **narrativa**), ele identificou a classe **Registro de relatos 1.3.2**.

No entanto, há alguns requisitos a serem cumpridospara que o programa possa ser usado de forma eficiente. Alguns desses requisitos já foram

mencionados antes, como o de o nome das classes da ontologia apresentar o código de classificação apenas no fim. Outra questão importante está na maneira como a busca é feita: a forma como o termo é escrito deve ser idêntica à forma como foi criada a instância na ontologia, de maneira que se torna importante que a lista de termos contenha apenas formas autorizadas. A organização da lista e a quantidade de termos para cada classe também devem ser bem definidas, tendo em vista que classes diferentes não podem ter todos os termos iguais (pelo menos um deve ser diferente). Outro problema pode ocorrer caso as classes possuam quantidades desiguais de termos: uma classe com menos termos pode apresentar em sua constituição os mesmos termos de outra classe com mais termos, de modo que aquela classe nunca seria o resultado final e único de uma busca (que é o que se espera obter).

Vale, por fim, destacar que o ideal é que o número de termos autorizados para busca seja grande e que estes sejam bastante abrangentes e gerais, sendo até recomendável aplicar uma pesquisa com toda a organização para selecionar todos os termos que poderiam ser usados.

Assim, é possível então afirmar que OntoClass alcança seu objetivo, porém cumpre ressaltar que, para obter resultados realmente positivos, deve-se obedecer a alguns requisitos relacionados especialmente à lista de termos autorizados, dentre os quais destacamos:

a) planos de classificação bem fundamentados em estudos sobre a organização e seus documentos;
b) organização da lista de termos autorizados com base nas características de trabalho da própria instituição, visto que essa lista deve ser de fácil compreensão e aplicação por aqueles responsáveis pela classificação dos documentos;
c) seleção de um mesmo número de termos para cada classe do plano de classificação, visando a evitar problemas oriundos de classes mais especificadas que outras, o que permitiria a existência de classes nunca dadas como único resultado possível de uma busca.

Os pontos acima, portanto, são fundamentais para uma boa utilização da versão atual do OntoClass. Evidentemente, estudos ainda poderão ser realizados

a fim de encontrar maneiras de reduzir esses requisitos em versões posteriores do *software*.

5 Considerações finais

A classificação, sendo uma das bases da gestão arquivística, deve primar pela consistência e deve ser realizada de maneira eficiente, para que os conjuntos documentais sejam bem definidos e para que os documentos sejam recuperáveis. Conforme foi discutido nas primeiras seções, se há problemas na classificação, outras funções arquivísticas, especialmente a avaliação e a descrição, podem ser prejudicadas, impedindo que a recuperação das informações presentes nos arquivos ocorra de modo satisfatório ou até mesmo levando documentos que deveriam ser permanentemente guardados à eliminação. É por conta desse perigo que soluções para a otimização da atividade de classificação devem ser pensadas.

Entre as possíveis causas para erros de classificação, pode estar o fato de que essa atividade é, em parte, subjetiva, ou seja, diante de um documento, o responsável pela classificação pode ficar em dúvida sobre a qual classe esse documento realmente pertence, de maneira que uma interpretação errônea pode levar aos problemas descritos no parágrafo anterior.

Diante disso e com base em estudos anteriores, foi elaborado um *software* que, por meio da aplicação do plano de classificação em ontologias, poderia indicar a um usuário a classe mais adequada a um documento que se procura classificar, tendo como base termos autorizados selecionados a partir de aspectos mais superficiais dos documentos. Assim, o OntoClass foi criado de modo que as classes da ontologia correspondessem às classes do plano de classificação e que as instâncias que compõem as classes fossem os termos autorizados.

Os primeiros testes, como foi possível perceber ao longo do texto, foram positivos e o programa é funcional, já podendo ser utilizado em situações concretas. Porém, seu funcionamento apropriado ainda impõe uma série de requisitos, que podem também ser vistos como limitações, ficando clara a importância de mais pesquisas sobre o tópico aqui trabalhado. É evidente

também que, por mais que o programa já demonstre o funcionamento esperado em seu desenvolvimento, há a necessidade de aplicação do OntoClass a uma instituição real, tanto na prática cotidiana quanto na preparação da lista de termos autorizados, visando a testar se seu objetivo (reduzir a subjetividade e garantir maior consistência na classificação) é realmente alcançado.

Portanto, este trabalho demonstrou como a classificação pode ser facilitada e tornada mais consistente com o auxílio de ontologias e com a aplicação delas em um *software* apropriado. Assim, o OntoClass, em sua atual versão de teste, cumpre o objetivo aqui proposto, porém apresenta limitações. Faz-se necessária sua aplicação em um contexto real para verificar se essa aplicação traz, de fato, benefícios, mesmo tendo em conta os requisitos anteriormente apresentados. Também se deve frisar que, no momento, atuando apenas na classificação, o OntoClass atende mais à busca contextual e não à de documentos individuais. Apesar disso, seu desenvolvimento dá margem para um aprofundamento ainda maior dos estudos relativos à recuperação da informação arquivística, tendo em vista a importância das ontologias para a ideia de fazer-se uma classificação a partir de metadados incluídos diretamente no sistema em uso.

Referências

BARROS, T. H. B.; GOMES, D. L. Classification and Knowledge Organization Systems: ontologies and archival classification. *In*: INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION. **Challenges and Opportunities for Knowledge Organization in the Digital Age.** Porto: Ergon Verlag, 2018. v. 16, p. 103-111.

BELLOTTO, H. L. **Arquivística:** objetos, princípios e rumos. São Paulo: Associação dos Arquivistas de São Paulo, 2002.

CARLAN, E.; MEDEIROS, M. B. B. Sistemas de organização do conhecimento na visão da Ciência da Informação. **Revista Ibero-Americana de Ciência da Informação,** Brasília, v. 4, n. 2, p. 53-73, 2011.

COOK, T. Arquivologia e pós-modernismo: novas formulações para velhos conceitos. **Informação Arquivística,** Rio de Janeiro, v. 1, n. 1, p. 123-148, 2012.

GONÇALVES, J. **Como classificar e ordenar documentos de arquivo**. São Paulo: Arquivo do Estado, 1998.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge acquisition,** Amsterdam, v. 5, n. 2, p. 199-220, 1993.

GUARINO, N. Understanding, building and using ontologies. **International Journal of Human-Computer Studies,** Amsterdam, v. 46, n. 2-3, p. 293-313, 1997.

MATTHEWS, B. Semantic Web Technologies. **E-Learning,** Bristol, v. 6, n. 6, p. 1-19, 2005.

MUSEN, M. A. The Protégé project: a look back and a look forward. **AI Matters,** Palo Alto, v. 1, n. 4, p. 4-12, 2015.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101** a guide to create your first ontology. [*S.l.: s.n.*], 2001.

ROUSSEAU, J.; COUTURE, C. **Os fundamentos da disciplina arquivística**. Lisboa: Dom Quixote, 1998.

SCHÄFER, M. B.; LIMA, E. S. A classificação e a avaliação de documentos: análise de sua aplicação em um sistema de gestão de documentos arquivísticos digitais. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 17, n. 3, p. 137-154, 2012.

SCHELLENBERG, T. R. **Arquivos modernos**: princípios e técnicas. 6. ed. Rio de Janeiro: FGV, 2006.

SOUSA, R. T. B. Alguns apontamentos sobre a classificação de documentos de arquivo. **Brazilian Journal of Information Science**, Marília, v. 8, n. 1/2, p. 1-24, 2014.

SOUSA, R. T. B. A classificação como função matricial do que-fazer arquivísticos. *In*: SANTOS, V. B.; INNARELI, H. C.; SOUSA, R. T. B. (org.). **Arquivística:** temas contemporâneos: classificação, preservação digital, gestão do conhecimento. Brasília: SENAC, 2007.

SOUSA, R. T. B. Os princípios arquivísticos e o conceito de classificação. *In*: RODRIGUES, G. M.; LOPES, I. L. (org.). **Organização e representação do conhecimento**. Brasília: Thesaurus, 2003. p. 240-269.

VITAL, L. P.; CAFÉ, L. M. A. Ontologias e taxonomias: diferenças. **Perspectivas em Ciência da Informação**. Belo Horizonte, v. 16, n. 2, p. 115-130, 2011.

A proposal for a tool for records classification based on ontologies

Abstract: This article describes and demonstrates a tool developed to reduce the subjective aspect inherent to archival classification, making it more consistent. Taking into account that classification errors can affect several other archival functions, especially appraisal and description, we've developed a software that we called Ontological Classifier (OntoClass). This software, through the creation of an ontology based on the classification scheme of a corporate body, is able to determine the class to which a document belongs based on terms arranged in a list. The theoretical basis was carried out through a bibliographic research and the development of the tool was made using the Python 3.7 programming language and the SPARQL query language. We conclude that although OntoClass reaches its goal, it is still necessary to test it in real situations and there are some requirements that must be met in order to achieve the desired results.

Keywords: Archival Science. Classification. Ontology. Query language. SPARQL.

Recebido: 13/04/2019
Aceito: 18/08/2019

Como citar:

GOMES, Daniel Libonati; BARROS, Thiago Henrique Bragato; SOUSA, Renato Tarciso Barbosa de; SANTOS JUNIOR, Roberto Lopes dos. Proposta de uma ferramenta para classificação arquivística com base em ontologias. **Em Questão,** Porto Alegre, v. 26, n. 1, p. 351-374, jan/abr. 2020. DOI: http://doi.org/10.19132/1808-5245261.