VII Seminário FESPSP - “Juventude, trabalho e profissão: desafios para o futuro no tempo presente”.
28 de outubro a 1 de novembro de 2019
GT 02 - Informação e ambientes digitais

PERSPECTIVA PARA A INDEXAÇÃO AUTOMÁTICA E SEMIAUTOMÁTICA: ESTUDO DA FERRAMENTA SISA

Vanda de Jesus Araújo[1]
Cibele Araújo Camargo Marques dos Santos[2]

Resumo: A indexação é uma representação sucinta de conteúdo e resulta da procura pelos assuntos principais dos documentos, gerando palavras isoladas representativas, os descritores. A indexação automática (IA) é realizada por sistema automatizado que utiliza índice pré-definido e regras lógicas que excluem palavras desnecessárias. O *Automatic Indexing System for Scientif Articles* (SISA), seleciona os descritores atribuídos ao texto, compara-os com o vocabulário controlado de um sistema de informação, uma lista de palavras vazias e relaciona em lista os termos candidatos a descritores. Sistemas automáticos são utilizados em conjunto com a indexação manual, configurando a indexação semiautomática (ISA). O projeto descrito neste artigo teve como objetivo apresentar o resultado da IA e ISA possibilitada pelo programa SISA, utilizando um *corpus* formado por artigos científicos. Foram comparados os dois processos de indexação, visando a localizar as principais diferenças de resultados e a indicar caminhos para o processo. O *corpus* foi formado por 34 artigos sobre indexação da Base BRAPCI, publicados em 2017. Uma lista de termos foi gerada a partir do SISA e confrontada com a indexação manual (IM) realizada com o Vocabulário Controlado da USP. Nas

[1] Graduanda em Biblioteconomia na Escola de Comunicações e Artes da Universidade de São Paulo. E-mail: vanda.araujo@usp.br
[2] Doutora em Ciência da Informação pela Escola de Comunicações e Artes da Universidade de São Paulo. E-mail: cibeleac@usp.br

correspondências exatas dos termos da IM e da IA, obtiveram-se em média 2,82 termos coincidentes. Entretanto, considerando termos de composições sintagmáticas recuperados de forma separada, houve correspondência de 6,1, representando uma cobertura de 79,6%. Considerou-se que o sistema pode ser uma ferramenta de auxílio ao indexador humano, sendo mais eficaz na ISA. Ademais foram constatadas necessidades de correções, revisões terminológicas e atualizações no Vocabulário USP.

Palavras-chave: Indexação automática, Indexação semi-automática, Vocabulário controlado, Organização e Representação do Conhecimento.

INTRODUÇÃO

A Organização e Representação do Conhecimento no processo de indexação, que tem por base a análise documentária, permite compactar a informação contida nos diversos documentos existentes em prol de uma recuperação consistente, através da análise, síntese e representação de conteúdo destes, que passam a ser traduzidos em termos de uma linguagem documentária.

As linguagens documentárias consistem num aparato técnico e lógico que proporciona o encontro do usuário da informação com seu objeto de busca. Nesse sentido, a indexação, ao lado do resumo documentário, constitui-se numa operação documentária de extrema importância para a recuperação da informação: "... *é a parte mais importante da análise documentária e condiciona o valor de um sistema documentário*..." (CHAUMIER,1980, *apud* SANTOS, 2009, p.03), ou seja, inserida em um sistema de informação devidamente arquitetado, proporciona ao usuário uma pesquisa eficaz, relevante e mais rápida, favorecendo a produção constante dos novos conhecimentos.

A indexação é uma representação enxuta de conteúdo e resulta da procura pelos assuntos principais dos documentos, gerando palavras representativas, os *termos descritores*. Quanto aos descritores utilizados na representação pela linguagem documentária, segundo Lancaster (*apud* SILVA e FUJITA, 2004, p.135), *"Devem ser indexadas as ideias do autor do texto e não as palavras".* A indexação é,

pois, um desafio conceitual dado a sua importância na recuperação dos documentos.

Do conhecimento resultante dos estudos de indexação, muitos métodos surgiram. A indexação manual (IM) é a mais convencional, como definida por Cunha (2008, 194): “é a indicação das palavras que representam os temas tratados num texto após a sua leitura”, a qual é feita por um indexador humano, razão pela qual é também chamada de indexação intelectual.

A indexação automática (IA), por sua vez, como o próprio termo explicita, é a realizada por um sistema automatizado. Na maioria das vezes, esse sistema utiliza um índice pré-definido, como por exemplo, um tesauro, e uma série de regras lógicas que excluem palavras desnecessárias, como é o caso de listas de palavras vazias, que contêm preposições, conjunções etc. Atualmente este sistema se desenvolve na esteira do aprimoramento computacional.

O *Automatic Indexing System for Scientif Articles* (SISA), em desenvolvimento pelo Professor Isidoro Gil Leiva da Universidade de Múrcia, é capaz de selecionar automaticamente os descritores atribuídos ao texto, compará-los com o vocabulário controlado de um sistema de informação e com uma lista de palavras consideradas vazias. Além disso, também relaciona, em lista à parte, os termos candidatos a descritores. No entanto, o SISA ainda se encontra em aperfeiçoamento e os estudos que estão sendo realizados em Ciência da Informação podem contribuir para que sua eficiência na representação da informação seja maior.

A IA pode ser considerada pouco acurada em relação à IM no que se refere ao refinamento semântico da indexação produzida pela mente humana. De fato, são poucos os sistemas de indexação puramente automáticos, o que reforça a necessidade de se estudarem suas lacunas e acertos.

A maioria dos sistemas automáticos são utilizados em conjunto com a indexação manual, configurando a indexação semiautomática (ISA), ou seja, um sistema que conta com etapas automáticas – geralmente as iniciais – mas que são refinadas por um indexador humano antes de se tornar produto documentário propriamente dito.

A consistência de indexação é um conceito-chave para a avaliação dos sistemas de indexação manual, automático e semiautomático, pois reflete o grau de correspondência de descritores entre indexações. Pela pesquisa de trabalhos já publicados, constatou-se que não há muitas publicações sobre o tema, mas Gil-

Leiva (1997) aponta que a consistência pode variar de 20 a 60%, o que representa uma margem de análise muito ampla. Portanto, ainda é difícil afirmar a superioridade qualitativa de algum dos três tipos de indexação mencionados.

Apesar da dificuldade apresentada acima, de modo geral, os autores lidos apresentam a indexação automática como recurso de grande potencial para facilitar a atividade indexadora. O que difere entre os autores é a forma com a qual um processo ainda tão desafiador, por tocar o funcionamento da linguagem e sua semântica, pode ser conduzido de modo a produzir uma indexação tal qual – ou melhor que – a feita por indexadores humanos.

Ao longo do tempo, vários critérios e pressupostos já foram adotados para aperfeiçoar a indexação. Um dos critérios estudados durante o projeto foi o da frequência média, que indica que os descritores tendem a estar numa faixa intermediária dos termos mais e menos frequentes. Sob essa perspectiva, um termo que ocorra muitas vezes em um determinado documento tende a ser geral demais para tornar a recuperação eficaz. Por outro lado, se um termo é muito pouco frequente, possivelmente não será capaz de sintetizar o conteúdo a se representar.

Outra forma de se estabelecerem regras que possam ser operacionalizadas pelos algoritmos a fim de selecionar descritores são as etiquetas morfológicas e sintáticas. No primeiro caso, há uma detecção de elementos mórficos que poderão, por exemplo, excluir candidatos a termo descritor, como é o caso das desinências em alguns contextos. No segundo, detectam-se elementos da própria estrutura dos períodos, como o emprego de sinais de pontuação.

Ademais foi estudado o uso dos elementos textuais dos artigos científicos, isto é, das diferentes partes que compõem o texto. Tais elementos são abordados de formas distintas pelos autores, porém, costumam-se atribuir diferentes valorações, que são maiores para título, resumo e texto, considerando-se que o título pode fornecer descritores com mais frequência e, se aliado ao resumo, pode, segundo alguns autores, apresentar exaustividade e precisão muito próxima à do texto completo do ponto de vista da indexação.

Como regra de eliminação, a lista de palavras vazias utilizadas pelos sistemas automáticos costuma ser muito útil, uma vez que consegue excluir até 50% dos candidatos a termo. Utilizando-se dessa lista, o sistema deixa de analisar elementos de certas classes de palavras, como preposições e conjunções, que só fazem

atrasar o processo de indexação pelo aumento de volume de palavras a serem processadas.

A precisão dos métodos automáticos de indexação só poderá aumentar diante de estudos que avaliem e indiquem possíveis saídas para os desenvolvedores de sistemas, acrescentando o olhar da Organização e Representação do Conhecimento, proposto por este projeto. Por sua vez, o trabalho do indexador tem muito a ganhar com um sistema de indexação automático ou semiautomático mais eficiente, pois as facilidades computacionais podem agilizar o processo de representação da informação, reduzir gastos e diminuir as incoerências conceituais humanas. Sem dúvida, a automatização da indexação constitui um intuito promissor para o futuro do campo científico relativamente à representação da produção de conhecimento.

De julho de 2018 a julho de 2019, desenvolvemos um projeto de iniciação científica embasado pelas considerações que até aqui foram discutidas. Este artigo tem como objetivo apresentar os resultados obtidos na pesquisa. Na ocasião, realizou-se uma comparação da IA e ISA possibilitada pelo programa SISA, utilizando um *corpus* formado por artigos científicos da área de indexação.

O objetivo primordial deste artigo é a análise das principais diferenças de resultados produzidas pelas respectivas indexações e indicar possíveis caminhos para uma ISA eficiente, a qual, desde que semanticamente satisfatória, representaria uma enorme vantagem no fator tempo, dinheiro e recursos humanos em relação à IM.

Inicialmente, foi feito um levantamento bibliográfico de artigos relacionados à indexação, a fim de se definir, no âmbito do projeto, o termo “indexação automática”. Assim, passamos a considerar que indexação é um conjunto de operações documentárias com o intuito de produzir a representação de um documento por descritores, tornando sua recuperação mais rápida e eficaz. A indexação automática, desse modo, consiste na realização de tais operações com o auxílio de ferramentas computacionais que visam a aumentar a eficiência e eficácia do processo, talvez aumentando a qualidade e poupando tempo e recursos.

A partir do levantamento bibliográfico e dos fichamentos resultantes, foi selecionada a literatura que serviria para embasar o estudo. Em especial, a leitura da tese do autor do SISA (GIL-LEIVA, 1997), ferramenta que nos serviu, ora de meio, ora de objeto, levando em conta os objetivos do projeto, e permitiu uma melhor

contextualização do ambiente metodológico no qual o modelo para indexação automática foi desenvolvido. Já o artigo sobre o SISA de Gil-Leiva (2017) permitiu-nos o conhecimento do funcionamento da ferramenta computacional.

Conforme a metodologia adotada, analisamos um *corpus* formado por artigos científicos da área de estudos indexação, selecionados da base de dados da BRAPCI, publicados no ano de 2017. Pelo filtro de busca “Todos os campos” combinado com o filtro “2017”, resultaram 34 artigos, sendo que o único descritor pesquisado foi “indexação”.

Uma vez feito o *download* dos artigos em formato pdf, demos prosseguimento à pesquisa com a realização da IM, independente da automática. Neste primeiro tratamento, não foi utilizado um vocabulário controlado. Uma lista provisória de termos foi elaborada.

A segunda operação desse tratamento é chamada de tradução pelo UNISIST (1975, p. 3). No âmbito do projeto, consistiu na IM que teve por escopo o Vocabulário Controlado da USP. Para tanto, os termos selecionados no primeiro tratamento foram confrontados com os descritores que constam no Vocabulário, que é uma linguagem construída a partir de procedimentos terminológicos e documentários, pelos bibliotecários do sistema de bibliotecas da Universidade de São Paulo, com a participação de especialistas de todas as áreas do conhecimento abrangidas pelos seus descritores, presentes no acervo de todas as bibliotecas da universidade. Esse vocabulário é o maior do Brasil e conta com a qualidade de bons profissionais em sua implementação e manutenção. O processo de IM nos forneceu uma lista de termos.

A próxima operação consistiu na IA, que se iniciou com a preparação dos artigos, o que envolveu: conversão para o formato de texto *txt*, colocação de etiquetas textuais, as quais servem para delimitar os componentes textuais que foram valorados conforme os critérios específicos da ferramenta SISA (Figura 1), como formulados por Gil-Leiva (2017). Dessa forma, os arquivos ficaram prontos para o processamento automático. Ao final desse processo, obteve-se uma segunda lista de descritores, selecionados automaticamente.

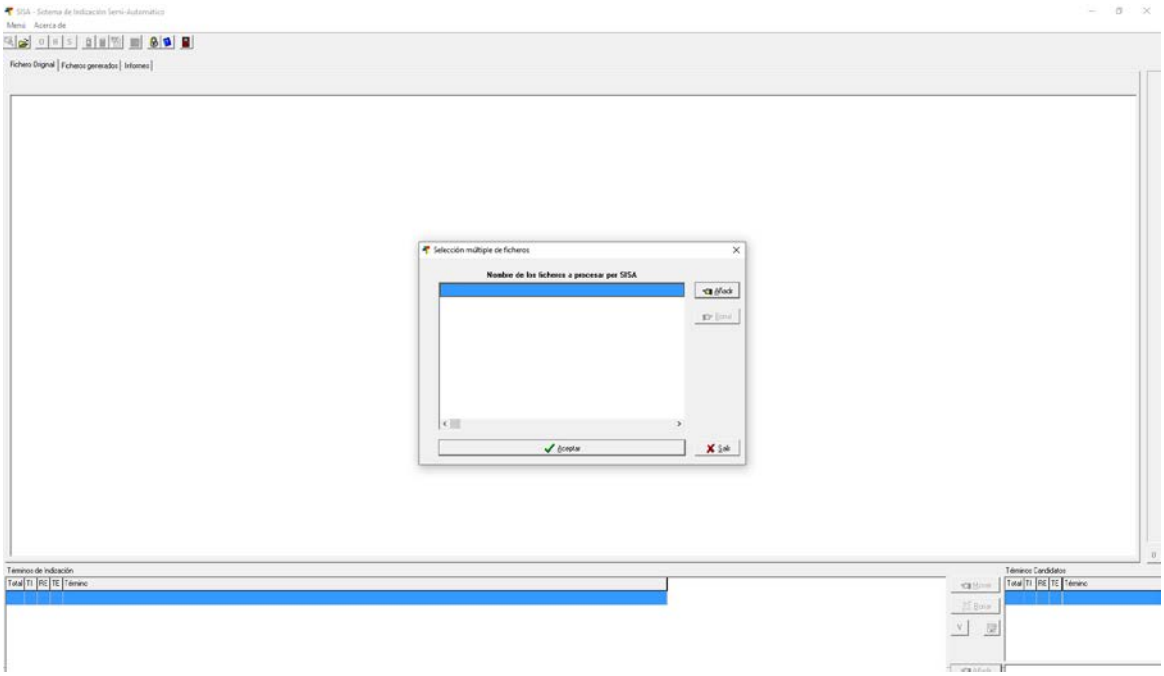


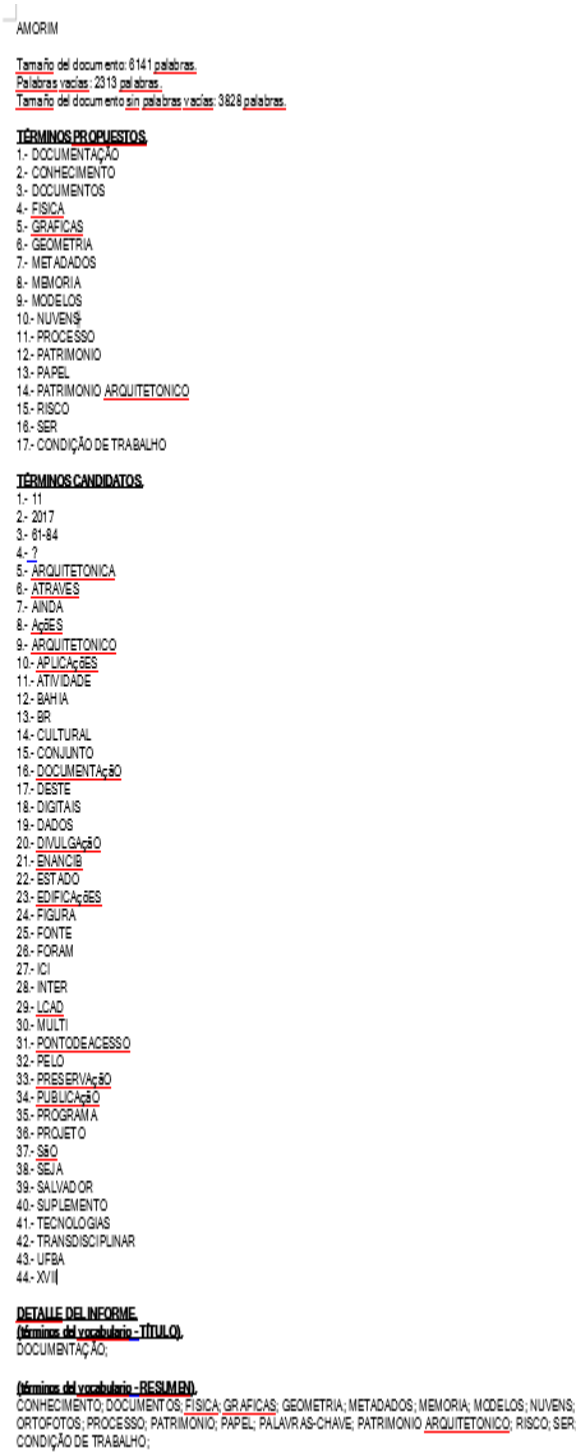
Figura 1 - Interface da ferramenta SISA

Fonte: captura de tela do SISA pela autora (2019)

Os termos resultantes das indexações foram analisados técnica e criticamente. A partir das conclusões obtidas, foi possível uma discussão acerca do tema e apresentação de uma terceira lista, de descritores propostos para a indexação semiautomática.

PROCEDIMENTOS E ANÁLISES

Para o entendimento da lista de termos gerada pelo SISA, é necessário diferenciar os termos que a ferramenta classifica como “termos candidatos” e “termos propostos” apresentados na Figura 2. Esses integram o resultado da IA propriamente dita; são os termos que “apareceriam” no caso de uma indexação puramente automática, pois a ferramenta identificou como relevantes, ao atribuir maior peso aos termos que ocorrem no título e resumo dos textos. Aqueles são termos que, devido ao número de ocorrências no *corpus*, poderiam ser úteis para o indexador humano realizar um trabalho mais eficaz, constituindo uma lista sugestiva para uma ISA.



AMORIM

Tamaño del documento: 6141 palabras.
Palabras vacías: 2313 palabras.
Tamaño del documento sin palabras vacías: 3828 palabras.

TÉRMINOS PROPUESTOS.
1.- DOCUMENTAÇÃO
2.- CONHECIMENTO
3.- DOCUMENTOS
4.- FISICA
5.- GRAFICAS
6.- GEOMETRIA
7.- METADADOS
8.- MEMORIA
9.- MODELOS
10.- NUVENS
11.- PROCESSO
12.- PATRIMONIO
13.- PAPEL
14.- PATRIMONIO ARQUITETONICO
15.- RISCO
16.- SER
17.- CONDIÇÃO DE TRABALHO

TÉRMINOS CANDIDATOS.
1.- 11
2.- 2017
3.- 61-84
4.- ?
5.- ARQUITETONICA
6.- ATRAVES
7.- AINDA
8.- AçõES
9.- ARQUITETONICO
10.- APLICAçõES
11.- ATIVIDADE
12.- BAHIA
13.- BR
14.- CULTURAL
15.- CONJUNTO
16.- DOCUMENTAçãO
17.- DESTE
18.- DIGITAIS
19.- DADOS
20.- DIVULGAçãO
21.- ENANCIB
22.- ESTADO
23.- EDIFICAçõES
24.- FIGURA
25.- FONTE
26.- FORAM
27.- ICI
28.- INTER
29.- LCAD
30.- MULTI
31.- PONTODEACESSO
32.- PELO
33.- PRESERVAçãO
34.- PUBLICAçãO
35.- PROGRAMA
36.- PROJETO
37.- SãO
38.- SEJA
39.- SALVADOR
40.- SUPLEMENTO
41.- TECNOLOGIAS
42.- TRANSDISCIPLINAR
43.- UFBA
44.- XVII

DETALLE DEL INFORME.
(términos del vocabulario - TÍTULO).
DOCUMENTAÇÃO;

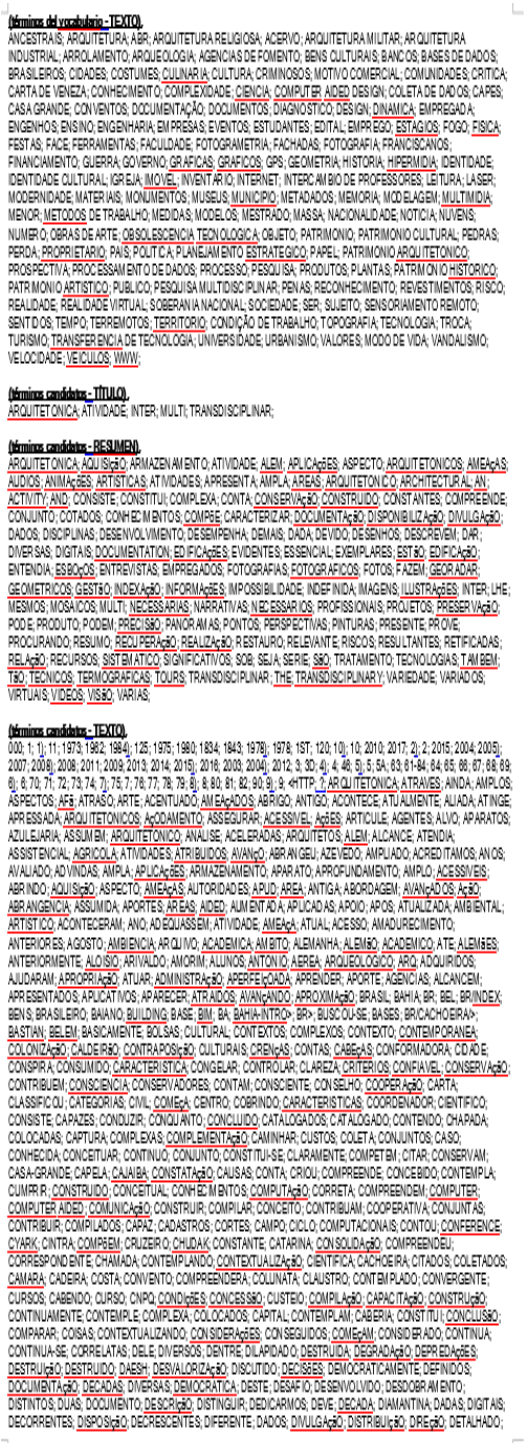
(términos del vocabulario - RESUMEN).
CONHECIMENTO; DOCUMENTOS; FISICA; GRAFICAS; GEOMETRIA; METADADOS; MEMORIA; MODELOS; NUVENS; ORTOFOTOS; PROCESSO; PATRIMONIO; PAPEL; PALAVRAS-CHAVE; PATRIMONIO ARQUITETONICO; RISCO; SER; CONDIÇÃO DE TRABALHO;

Figura 2 - Exemplo de listas produzidas pelo SISA.

Fonte: Elaborado pela autora (2019) a partir do SISA

O sistema gera outras três listas: “termos do vocabulário - TITULO”, “termos do vocabulário - RESUMO” e “termos do vocabulário - TEXTO” (Figura 3), que

contém todas as palavras do artigo que não foram eliminadas pela lista de palavras vazias.



(Número del vocabulario - TEXTO).
ANCESTRAIS; ARQUITETURA; ABR; ARQUITETURA RELIGIOSA; ACERVO; ARQUITETURA MILITAR; ARQUITETURA INDUSTRIAL; ARROLAMENTO; ARQUEOLOGIA; AGENCIAS DE FOMENTO; BENS CULTURAIS; BANCOS; BASES DE DADOS; BRASILEIROS; CIDADES; COSTUMES; CULINARIA; CULTURA; CRIMINOSOS; MOTIVO COMERCIAL; COMUNIDADES; CRITICA; CARTA DE VENEZA; CONHECIMENTO; COMPLEXIDADE; CIENCIA; COMPUTER AIDED DESIGN; COLETA DE DADOS; CAPES; CASA GRANDE; CONVENTOS; DOCUMENTAÇÃO; DOCUMENTOS; DIAGNOSTICO; DESIGN; DINAMICA; EMPREGADA; ENGENHOS; ENSINO; ENGENHARIA; EMPRESAS; EVENTOS; ESTUDANTES; EDITAL; EMPREGO; ESTAGIOS; FOGO; FISICA; FESTAS; FACE; FERRAMENTAS; FACULDADE; FOTOGRAMETRIA; FACHADAS; FOTOGRAFIA; FRANCISCANOS; FINANCIAMENTO; GUERRA; GOVERNO; GRAFICAS; GRAFICOS; GPS; GEOMETRIA; HISTORIA; HIPERMIDIA; IDENTIDADE; IDENTIDADE CULTURAL; IGREJA; IMOVEL; INVENTARIO; INTERNET; INTERCAMBIO DE PROFESSORES; LEITURA; LASER; MODERNIDADE; MATERIAIS; MONUMENTOS; MUSEUS; MUNICIPIO; METADADOS; MEMORIA; MODELAGEM; MULTIMIDIA; MENOR; METODOS DE TRABALHO; MEDIDAS; MODELOS; MESTRADO; MASSA; NACIONALIDADE; NOTICIA; NUVENS; NUMERO; OBRAS DE ARTE; OBSOLESCENCIA TECNOLOGICA; OBJETO; PATRIMONIO; PATRIMONIO CULTURAL; PEDRAS; PERDA; PROPRIETARIO; PAIS; POLITICA; PLANEJAMENTO ESTRATEGICO; PAPEL; PATRIMONIO ARQUITETONICO; PROSPECTIVA; PROCESSAMENTO DE DADOS; PROCESSO; PESQUISA; PRODUTOS; PLANTAS; PATRIMONIO HISTORICO; PATRIMONIO ARTISTICO; PUBLICO; PESQUISA MULTIDISCIPLINAR; PENAS; RECONHECIMENTO; REVESTIMENTOS; RISCO; REALIDADE; REALIDADE VIRTUAL; SOBERANIA NACIONAL; SOCIEDADE; SER; SUJEITO; SENSORIAMENTO REMOTO; SENTIDOS; TEMPO; TERREMOTOS; TERRITORIO; CONDIÇÃO DE TRABALHO; TOPOGRAFIA; TECNOLOGIA; TROCA; TURISMO; TRANSFERENCIA DE TECNOLOGIA; UNIVERSIDADE; URBANISMO; VALORES; MODO DE VIDA; VANDALISMO; VELOCIDADE; VEICULOS; WWW;

(Número candidatos - TÍTULO).
ARQUITETONICA; ATIVIDADE; INTER; MULTI; TRANSDISCIPLINAR;

(Número candidatos - RESUMEN).
ARQUITETONICA; AQUISIçãO; ARMAZENAMENTO; ATIVIDADE; ALEM; APLICAçõES; ASPECTO; ARQUITETONICOS; AMEAçAS; AUDIOS; ANIMAçõES; ARTISTICAS; ATIVIDADES; APRESENTA; AMPLA; AREAS; ARQUITETONICO; ARCHITECTURAL; AN; ACTIVITY; AND; CONSISTE; CONSTITUI; COMPLEXA; CONTA; CONSERVAçãO; CONSTRUIDO; CONSTANTES; COMPREENDE; CONJUNTO; COTADOS; CONHECIMENTOS; COMPRE; CARACTERIZAR; DOCUMENTAçãO; DISPONIBILIZAçãO; DIVULGAçãO; DADOS; DISCIPLINAS; DESENVOLVIMENTO; DESEMPENHA; DEMAIS; DADA; DEVIDO; DESENHOS; DESCREVEM; DAR; DIVERSAS; DIGITAIS; DOCUMENTATION; EDIFICAçõES; EVIDENTES; ESSENCIAL; EXEMPLARES; ESTãO; EDIFICAçãO; ENTENDIA; ESBOçOS; ENTREVISTAS; EMPREGADOS; FOTOGRAFIAS; FOTOGRAFICOS; FOTOS; FAZEM; GEORADAR; GEOMETRICOS; GESTãO; INDEXAçãO; INFORMAçõES; IMPOSSIBILIDADE; INDEFINIDA; IMAGENS; ILUSTRAçõES; INTER; LHE; MESMOS; MOSAICOS; MULTI; NECESSARIAS; NARRATIVAS; NECESSARIOS; PROFISSIONAIS; PROJETOS; PRESERVAçãO; PODE; PRODUTO; PODEM; PRECISãO; PANORAMAS; PONTOS; PERSPECTIVAS; PINTURAS; PRESENTE; PROVE; PROCURANDO; RESUMO; RECUPERAçãO; REALIZAçãO; RESTAURO; RELEVANTE; RISCOS; RESULTANTES; RETIFICADAS; RELAçãO; RECURSOS; SISTEMATICO; SIGNIFICATIVOS; SOB; SEJA; SERIE; SãO; TRATAMENTO; TECNOLOGIAS; TAMBEM; TãO; TECNICOS; TERMOGRAFICAS; TOURS; TRANSDISCIPLINAR; THE; TRANSDISCIPLINARY; VARIEDADE; VARIADOS; VIRTUAIS; VIDEOS; VISãO; VARIAS;

(Número candidatos - TEXTO).
000; 1; 1); 11; 1973; 1982; 1984); 125; 1975; 1980; 1834; 1843; 1978); 1978; 1ST; 120; 10); 10; 2010; 2017, 2); 2; 2015; 2004; 2005); 2007; 2008); 2008; 2011; 2009; 2013; 2014, 2015); 2016; 2003; 2004); 2012; 3; 3D; 4); 4; 46; 5); 5; 5A; 63; 61-84; 64; 65; 66; 67; 68; 69; 6); 6; 70; 71; 72; 73; 74; 7); 75; 7; 76; 77; 78; 79; 8); 8; 80; 81; 82; 90; 9); 9; <HTTP; ?; ARQUITETONICA; ATRAVES; AINDA; AMPLOS; ASPECTOS; AFS; ATRASO; ARTE; ACENTUADO; AMEAçADOS; ABRIGO; ANTIGO; ACONTECE; ATUALMENTE; ALIADA; ATINGE; APRESSADA; ARQUITETONICOS; AçODAMENTO; ASSEGURAR; ACESSIVEL; AçõES; ARTICULE; AGENTES; ALVO; APARATOS; AZULEJARIA; ASSUMBI; ARQUITETONICO; ANALISE; ACELERADAS; ARQUITETOS; ALEM; ALCANCE; ATENDIA; ASSISTENCIAL; AGRICOLA; ATIVIDADES; ATRIBUIDOS; AVANçO; ABRANGEU; AZEVEDO; AMPUADO; ACREDITAMOS; ANOS; AVALIADO; ADVINDAS; AMPLA; APLICAçõES; ARMAZENAMENTO; APARATO; APROFUNDAMENTO; AMPLO; ACESSIVEIS; ABRINDO; AQUISIçãO; ASPECTO; AMEAçAS; AUTORIDADES; APUD; AREA; ANTIGA; ABORDAGEM; AVANçADOS; Açãº; ABRANGENCIA; ASSUMIDA; APORTES; AREAS; AIDED; AUMENTADA; APLICADAS; APOIO; APOS; ATUALIZADA; AMBIENTAL; ARTISTICO; ACONTECERAM; ANO; ADEQUASSEM; ATIVIDADE; AMEAçA; ATUAL; ACESSO; AMADURECIMENTO; ANTERIORES; AGOSTO; AMBIENCIA; ARQUIVO; ACADEMICA; AMBITO; ALEMANHA; ALEMãO; ACADEMICO; ATE; ALEMãES; ANTERIORMENTE; ALOISIO; ARIVALDO; AMORIM; ALUNOS; ANTONIO; AEREA; ARQUEOLOGICO; ARQ; ADQUIRIDOS; AJUDARAM; APROPRIAçãO; ATUAR; ADMINISTRAçãO; APERFEIçOADA; APRENDER; APORTE; AGENCIAS; ALCANCEM; APRESENTADOS; APLICATIVOS; APARECER; ATRAIDOS; AVANçANDO; APROXIMAçãO; BRASIL; BAHIA; BR; BEL; BR/INDEX; BENS; BRASILEIRO; BAIANO; BUILDING; BASE; BIM; BA; BAHIA-INTRO>; BR>; BUSCOU-SE; BASES; BR/CACHOEIRA>; BASTIAN; BELEM; BASICAMENTE; BOLSAS; CULTURAL; CONTEXTOS; COMPLEXOS; CONTEXTO; CONTEMPORANEA; COLONIZAçãO; CALDEIRãO; CONTRAPOSIçãO; CULTURAIS; CRENçAS; CONTAS; CABEçAS; CONFORMADORA; CIDADE; CONSPIRA; CONSUMIDO; CARACTERISTICA; CONGELAR; CONTROLAR; CLAREZA; CRITERIOS; CONFIAVEL; CONSERVAçãO; CONTRIBUEM; CONSCIENCIA; CONSERVADORES; CONTAM; CONSCIENTE; CONSELHO; COOPERAçãO; CARTA; CLASSIFICOU; CATEGORIAS; CIVIL; COMEçA; CENTRO; COBRINDO; CARACTERISTICAS; COORDENADOR; CIENTIFICO; CONSISTE; CAPAZES; CONDUZIR; CONJUNTO; CONCLUIDO; CATALOGADOS; CATALOGADO; CONTENDO; CHAPADA; COLOCADAS; CAPTURA; COMPLEXAS; COMPLEMENTAçãO; CAMINHAR; CUSTOS; COLETA; CONJUNTOS; CASO; CONHECIDA; CONCEITUAR; CONTINUO; CONJUNTO; CONSTITUI-SE; CLARAMENTE; COMPETEM; CITAR; CONSERVAM; CASA-GRANDE; CAPELA; CAJAIBA; CONSTATAçãO; CAUSAS; CONTA; CRIOU; COMPREENDE; CONCEBIDO; CONTEMPLA; CUMPRIR; CONSTRUIDO; CONCEITUAL; CONHECIMENTOS; COMPUTAçãO; CORRETA; COMPREENDEM; COMPUTER; COMPUTER AIDED; COMUNICAçãO; CONSTRUIR; COMPILAR; CONCEITO; CONTRIBUAM; COOPERATIVA; CONJUNTAS; CONTRIBUIR; COMPILADOS; CAPAZ; CADASTROS; CORTES; CAMPO; CICLO; COMPUTACIONAIS; CONTOU; CONFERENCE; CYARK; CINTRA; COMPREM; CRUZEIRO; CHUDAK; CONSTANTE; CATARINA; CONSOLIDAçãO; COMPREENDEU; CORRESPONDENTE; CHAMADA; CONTEMPLANDO; CONTEXTUALIZAçãO; CIENTIFICA; CACHOEIRA; CITADOS; COLETADOS; CAMARA; CADEIRA; COSTA; CONVENTO; COMPREENDERA; COLUNATA; CLAUSTRO; CONTEMPLADO; CONVERGENTE; CURSOS; CABENDO; CURSO; CNPQ; CONDIçõES; CONCESSãO; CUSTEIO; COMPILAçãO; CAPACITAçãO; CONSTRUçãO; CONTINUAMENTE; CONTEMPLE; COMPLEXA; COLOCADOS; CAPITAL; CONTEMPLAM; CABEIRA; CONSTITUI; CONCLUSãO; COMPARAR; COISAS; CONTEXTUALIZANDO; CONSIDERAçõES; CONSEGUIDOS; COMEçAM; CONSIDERADO; CONTINUA; CONTINUA-SE; CORRELATAS; DELE; DIVERSOS; DENTRE; DILAPIDADO; DESTRUIDA; DEGRADAçãO; DEPREDAçõES; DESTRUIçãO; DESTRUIDO; DAESH; DESVALORIZAçãO; DISCUTIDO; DECISõES; DEMOCRATICAMENTE; DEFINIDOS; DOCUMENTAçãO; DECADAS; DIVERSAS; DEMOCRATICA; DESTE; DESAFIO; DESENVOLVIDO; DESDOBRAMENTO; DISTINTOS; DUAS; DOCUMENTO; DESCRIçãO; DISTINGUIR; DEDICARMOS; DEVE; DECADA; DIAMANTINA; DADAS; DIGITAIS; DECORRENTES; DISPOSIçãO; DECRESCENTES; DIFERENTE; DADOS; DIVULGAçãO; DISTRIBUIçãO; DRIçãO; DETALHADO;

Figura 3 – Listas completas produzidas pelo SISA.

Fonte: Elaborado pela autora (2019) a partir do SISA

Apresenta-se no Quadro 1 o comparativo das indexações para o artigo 1 do *corpus* analisado a fim de retratar com mais clareza as diferenças e semelhanças encontradas. O mesmo procedimento foi adotado para cada um dos 34 artigos selecionados.

Artigo 1

AMORIM, A. L. A documentação arquitetônica como uma atividade multi, inter e transdisciplinar. Ponto de Acesso, v. 11, n. 1, 201710.9771/rpa.v11i1.23176. Disponível em: <http://www.brapci.inf.br/v/a/23508>. Acesso em: 19 Maio 2018.

Indexação Manual (IM)	**Indexação automática (IA)**	**Indexação semiautomática (ISA)**
DOCUMENTACAO	DOCUMENTACAO	DOCUMENTACAO
INTERDISCIPLINARIDADE	INTER* TRANSDISCIPLINAR* MULTI* PESQUISA MULTIDISCIPLINAR**	INTERDISCIPLINARIDADE TRANSDISCIPLINARIDADE MULTIDISCIPLINARIDADE
TECNOLOGIA DA INFORMACAO	TECNOLOGIAS* INFORMACAO*	TECNOLOGIA DA INFORMACAO
PATRIMONIO ARQUITETONICO	PATRIMONIO ARQUITETONICO	PATRIMONIO ARQUITETONICO
MEMORIA/PRESERVACAO	MEMORIA PRESERVACAO*	MEMORIA/PRESERVACAO
TRATAMENTO DA INFORMACAO	TRATAMENTO INFORMACAO**	TRATAMENTO DA INFORMACAO
PATRIMONIO ARQUITETONICO/ PRESERVACAO	RISCO PATRIMONIO ARQUITETONICO PRESERVACAO*	PATRIMONIO ARQUITETONICO/ PRESERVACAO
	CONDICAO DE TRABALHO	
	PROCESSO	
	FISICA	
	PAPEL	

	METADADOS	
	MODELOS	
	NUVENS	
	GRAFICAS	
	GEOMETRIA	
	DOCUMENTOS	
	CONHECIMENTO	
	SER	

Quadro 1 - Indexações do artigo 1

Fonte: elaborado pela autora (2019)

Termos que figuram como “termos candidatos” e não como “termos propostos” estão indicados no Quadro 1 com asterisco (*).

Já termos que constam apenas na lista de “termos do vocabulário - TEXTO” ou “termos do vocabulário - RESUMO” estão indicados com dois asteriscos (**). Esta última lista, porém, é praticamente exaustiva e significaria grande trabalho se o indexador tivesse que consultá-la a cada indexação.

A primeira coluna, “Indexação Manual”, elenca o resultado dessa indexação já confrontado com o Vocabulário USP. A constatação mais evidente durante a primeira operação de definição de descritores revelou que muitos dos termos procurados no vocabulário controlado não foram encontrados.

São exemplos disso os termos: taxonomia, informação digital, repositórios digitais, classificação bibliográfica, classificação facetada, organização da informação, plataformas digitais, sistemas de informação, mapa conceitual, RDF, necessidade informacional, expressão de busca, revocação, documentos iconográficos, *linked data*, tecnologias digitais.

A segunda coluna elenca os termos candidatos e propostos de forma automática pelo SISA. Organizamos os termos de modo a ficarem na linha mais próxima semanticamente do termo da IM.

A terceira coluna traz uma proposta de ISA, resultado da comparação das duas primeiras colunas e da análise das mesmas, que são explicitadas nos comentários apresentados na sequência de cada quadro analítico, como indicado no Quadro 2.

Comentário: Inicialmente o termo "INTERDISCIPLINARIDADE" foi considerado de extrema importância na indexação manual, já que parece ser o mais atual em termos pedagógicos para representar o conceito "atividade multi, inter e transdisciplinar", presente no título. Entretanto não poderia figurar inteiro, isto é, com todas as letras, por não aparecer no artigo desse modo. Pode-se depreender, então, que a IM foi capaz de atribuir um descritor importante, que a IA não conseguiria *per si*. Após breve pesquisa, porém, achamos melhor, afinal, incluir os descritores INTERDISCIPLINARIDADE; TRANSDISCIPLINARIDADE – por influência do candidato da IA TRANSDISCIPLINAR e MULTIDISCIPLINARIDADE, já que a IA nos forneceu PESQUISA MULTIDISCIPLINAR. A complementaridade entre IA e IM aumentou, portanto o número de descritores da indexação final, semiautomática. Figuram como termos propostos os prefixos "INTER" e "MULTI", os quais, na verdade, não compõem o Vocabulário USP. Talvez essa ocorrência inesperada se explique pelo tipo de composição (prefixação) que apresentam, isto é, esses são prefixos frequentemente seguidos por hífen em nossa língua.

Quadro 2 - Análise da indexações do artigo 1
Fonte: elaborado pela autora (2019)

Em relação às análises, para efeito da pesquisa, tanto nos termos do vocabulário USP quanto nos termos do SISA não foram empregados acentos ou sinais gráficos nos termos.

RESULTADOS E DISCUSSÃO

O primeiro olhar lançado aos quadros elaborados buscou as correspondências exatas dos termos da IM e da IA. Consideramos como exata a incidência das mesmas letras, desprezando-se a especificação de área que aparece entre parênteses no vocabulário USP, como no exemplo: INDEXACAO (BIBLIOTECONOMIA). De fato, esse tipo de correspondência foi pouco significativa,

se considerarmos apenas que, em média, obtiveram-se 2,82 termos coincidentes por artigo e que os números de correspondências variaram entre 0 e 8.

Todavia, verificou-se uma tendência da IA em apontar termos simples, isto é, não listar composições sintagmáticas, como ocorre com TECNOLOGIA DA INFORMACAO E TRATAMENTO DA INFORMACAO. Nesses casos, recuperou-se cada elemento em separado (desmembrado), sendo que a maioria eram substantivos (por exemplo: TECNOLOGIA e INFORMACAO, TRATAMENTO e INFORMACAO), ou recuperou-se apenas um dos termos (por exemplo: apenas INFORMACAO), lembrando o *uniterm* mencionado por Lancaster (2004, p.19), forma de indexação, se bem que derivada, que empregava apenas termos formados por uma única palavra para representar o conteúdo temático.

Se considerarmos como correspondência também esses termos desmembrados, isto é, recuperados em separado pela IA e também desprezarmos alguns detalhes desinenciais da língua, como é o caso da variação das terminações nominais em gênero e número, teremos um número de termos correspondentes por artigo muito mais significativo: 6,1.

Artigo	Correspondências exatas	Correspondências com desmembramentos	Termos da ISA
1	2	6	7
2	1	4	6
3	2	4	7
4	2	7	10
5	2	4	7
6	5	6	8
7	2	5	5
8	4	5	9

9	2	7	7
10	4	6	7
11	5	5	6
12	2	3	5
13	3	6	10
14	4	8	10
15	2	9	10
16	1	5	6
17	3	5	9
18	3	6	6
19	1	5	8
20	2	7	7
21	8	10	11
22	1	6	7
23	2	7	7
24	4	4	5
25	1	9	9
26	5	8	8
27	0	6	6
28	1	6	8
29	2	4	7
30	4	4	8

31	4	6	9
32	4	5	6
33	4	11	11
34	4	9	9
TOTAL	96	208	261

Quadro 2- Correspondências de termos das indexações manuais e automáticas
Fonte: Elaborado pela autora (2019)

Sob esse prisma, a “cobertura temática” da IA é muito mais eficaz, o que pode ser comprovado pelo número total de termos gerados por ela (208) em relação ao número total da ISA (261), representando uma cobertura de 79,6%. A cobertura total também reforça a eficiência que poderia ser atribuída à IA do SISA, no caso das indexações de 9 artigos e, nos casos em que a cobertura não foi total, foi aproximada.

Apesar dessa eficiência aparente, um fato metodológico pode enfraquecer a confiança numa indexação puramente automática: a segunda coluna que apresentamos na análise de cada artigo, numerado de 1 a 34 (Quadro 2), trazia, além dos descritores da lista de termos de indexação, que, como dissemos, são os termos da IA propriamente dita, também termos candidatos, listados em lista à parte pelo SISA, marcados com um asterisco (*) e termos de outra lista, exaustiva na prática, não considerada pelo sistema como descritores, mas listados por ocorrência, marcados com dois asteriscos (**). O que se pode depreender dessas informações é que esse grau de cobertura da IA deve parte de seu alcance ao esforço intelectual da indexadora. Em outras palavras, a IA fornece as “pistas”, mas a construção de uma lista de descritores mais completa dependeu da intuição, humana, e não da automatização do processo.

Uma possibilidade de interpretação para a segmentação operada pelo SISA são as implicações da exclusão de termos vazios. Esses elementos são indicados em lista apropriada a ser utilizada pelo SISA e, geralmente, pertencem à classe

gramatical das preposições, conjunções e verbos. Pode ser que a desconsideração de uma preposição como "*de"* provoque o desmembramento de um termo como TRATAMENTO DA INFORMACAO. Isso seria então um problema a ser resolvido pelo programador. Os termos (sintagmáticos) como ARTIGOS DE PERIÓDICOS ou TRIBUNAL DE CONTAS*,* porém, ocorreram juntos, incluindo as preposições, na IA, o que enfraquece a hipótese há pouco levantada.

Não identificamos a regra lógica para tal "escolha" do SISA, uma vez que, por exemplo, os termos USUARIOS e INFORMAÇAO (que sempre aparecem desmembrados) compõem o vocabulário USP do modo composto USUARIOS DA INFORMACAO.

Os termos CIENCIA e INFORMACAO sempre figuram como termos separados na IA do SISA, isto é, não houve a ocorrência de CIENCIA DA INFORMACAO, embora esse termo componha o Vocabulário USP e seja abundante nos artigos. Levantamos a hipótese de que o termo Ciência da Informação talvez se encontre apenas desmembrado por não aparecer com tanta frequência no título e resumo dos artigos, elementos pré-textuais de maior relevância para a ferramenta computacional utilizada.

O segundo olhar lançado considerou as vantagens da IM por possibilitar a atribuição de um termo subentendido na leitura do artigo, mas não explícito. Alguns exemplos:

- INVESTIGACAO CRIMINAL - não havia no texto do artigo a palavra "criminal"
- MOTORES DE BUSCA – não constava no texto analisado, mas o contexto permitia indicar o termo.
- FONTES DE INFORMACAO – conceito não descrito, mas implícito em um dos noartigo artigo analisado.
- JOGOS EDUCATIVOS - não ocorre no artigo, mas foi uma alternativa para o termo candidato GAMIFICAÇÃO
- GAMIFICACAO - não consta no vocabulário controlado, portanto, não poderia ser indicado pela IA

Talvez possamos afirmar que a IM possibilita que conceitos que não integram o texto possam ser indicados como descritores. E isso a IA, pela carência semântica, não é capaz de realizar ainda.

Termo-chave de busca para os artigos selecionados, INDEXACAO não figurou entre os termos propostos da IA nenhuma vez. Ocorreu apenas como termo candidato (CANDIDATO*), ou como simples termo do vocabulário (TEXTO ou RESUMO), o que pode significar que talvez o fato de o termo no vocabulário controlado estar apresentado na forma INDEXACAO (BIBLIOTECONOMIA), com parênteses e especificação da área, para ser diferenciado da indexação de outras áreas de conhecimento, possa ter tido impedida a sua maior valoração pela IA.

A palavra SER apareceu em 8 indexações como termo proposto, representando um número significativo de ocorrências nesta pesquisa. Sendo um verbo, poderia ter sido incluído na lista de palavras vazias. Entretanto a palavra “ser” pode pertencer à classe gramatical dos substantivos e, portanto, compor o vocabulário controlado. Como o SISA ainda não é capaz de realizar a desambiguação semântica, mais uma vez podemos afirmar que a IM foi crucial para determinar quando a palavra “ser” era vazia, ou seja, um não-termo.

Apesar dos pontos que precisam ser aprimorados nessa ferrementa ainda em desenvolvimento, uma importante contribuição da IA para a indexação final, a ISA, foram os termos que, sem a visualização da lista de candidatos a termos do sistema, poderiam passar despercebidas numa indexação puramente manual. Foi o que ocorreu em 12 dos 34 artigos analisados. Em todos esses casos, o SISA acrescentou descritores à ISA, contribuindo com uma melhor cobertura temática da indexação.

CONSIDERAÇÕES FINAIS

Apesar de Gil-Leiva (1997) descrever trabalhos acerca do tema abordado, tanto o número de indexações realizadas, quanto a metodologia adotada pelos pesquisadores eram muito diversificadas para despertar certezas mais concretas acerca da qualidade da IA relativamente à IM. Prova disso é que as pesquisas realizadas até agora não foram capazes de estimar o grau de consistência entre duas IA.

Na realidade, os artigos estudados pelo autor demonstram uma concentração de pesquisa em documentos de apenas algumas áreas do conhecimento, como as das ciências exatas e medicina, por apresentarem produção científica mais padronizada. Isso constituiria mais um fator limitante da abrangência para um estudo que se propõe observar a qualidade de um sistema IA. Essas questões ainda em aberto demonstram a necessidade da continuidade de pesquisas que abordem as diferentes formas de indexação, atentando para a consistência.

No que diz respeito ao comportamento do SISA, pudemos ver que tem se mostrado uma ferramenta de auxílio ao indexador humano, sendo mais eficaz quando num sistema de ISA. Há correções que talvez já estejam sendo sanadas numa atualização que está em curso pelo autor do sistema em versão web. Embora não possamos indicar as soluções computacionais para o problema, indicamos alguns dos empecilhos encontrados durante as indexações e esperamos que isso possa contribuir para a melhoria do sistema.

Ademais, constatamos a necessidade de correções, revisões terminológicas e atualizações no Vocabulário USP. Sendo o escopo conceitual para a IA, IM ou ISA, essa lista pré-definida de termos também deve estar em constante aperfeiçoamento, tendo sua qualidade garantida por recursos humanos e financeiros compatíveis com sua importância para a produção científica.

REFERÊNCIAS

AMORIM, A. L. A documentação arquitetônica como uma atividade multi, inter e transdisciplinar. Ponto de Acesso, v. 11, n. 1, 201710.9771/rpa.v11i1.23176. Disponível em: <http://www.brapci.inf.br/v/a/23508>. Acesso em: 19 Maio 2018.

CUNHA, Murilo Bastos da. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008.

GIL-LEIVA, Isidoro. SISA - Automatic Indexing System for Scientific Articles: experiments with location heuristics rules versus TF-IDF rules. **Knowledge Organization,** v.44, n.3, p.139-162, jan. 2017.

GIL-LEIVA, Isidoro. **La automatización de la indexación, propuesta teórico-metodológica: aplicación al área de biblioteconomia y documentación**. 1997. 268 f. Tese (Doutorado) - Departamento de Información y Documentación. Universidad de Murcia, Murcia. Disponível em: <https://webs.um.es/isgil/resources/PhDissertation%20Gil-Leiva.pdf>. Acesso em: 18 dez. 2018.

KOBASHI, N. Y. **Elaboração de informações documentárias:** em busca de uma metodologia. 1994. Tese (Doutorado em Ciência da Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 1994.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. 2 ed. Brasília: Briquet de Lemos/Livros, 2004.

SANTOS, V. N. **Indexação automática de documentos textuai**s: iniciativas dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72f . Trabalho de Conclusão de Curso (Bacharelado em Biblioteconomia) – Departamento de Biblioteconomia e Documentação, Escola de Comunicações e Artes da Universidade de São Paulo, 2009.

SILVA, M. R.; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, n.16, v. 2, p.133-161, maio/ago., 2004.

UNISIST. **Indexing principles**. Paris: UNESCO, 1976.