RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

‘Proveniência’ na terminografia arquivística de língua portuguesa: prospeção e visualização de (dis)similaridades em termos e definições

L. S. Ascensão de Macedo
Universidade de Coimbra, Centro de Estudos Interdisciplinares do Século XX, Núcleo Património e Humanidades Digitais, Coimbra, Portugal
ascensaodemacedo@gmail.com

DOI: https://doi.org/10.26512/rici.v11.n2.2018.8334

Resumo: Este artigo estuda o termo ‘proveniência’ com base na terminografia arquivística publicada em Portugal e no Brasil (1986-2013). Analisam-se ocorrências do termo ‘proveniência’ em entradas terminológicas e definições com o objetivo de identificar relações de similaridade/dissimilaridade em *datasets*. Adota-se uma abordagem semasiológica baseada em corpora diacrónicos com recurso a ferramentas de prospeção de dados textuais. Das 2760 entradas terminológicas, o termo ‘proveniência’ ocorre em 71, manifestando-se quer de forma monolexemática quer polilexemática. Verifica-se escassa reutilização de termos e definições na terminografia arquivística em contexto intralinguístico.

ARTIGOS

Palavras-chave: Arquivologia; Língua portuguesa; Princípio da proveniência; Terminologia.

‘Provenance’ in the Portuguese archival terminography: mining and visualizing similarities between terminological terms and definitions

Abstract: This paper focuses on the term ‘proveniência’ (provenance), based on the Portuguese archival terminography published in Portugal and Brazil (1986-2013). We analyze similarity/dissimilarity between terminological entries and definitions based on diachronic corpora. We used a semasiological approach using basic text mining techniques and text clustering methods with a dendrific visualization. The term ‘proveniência’ occurs in 71 (from 2760) entries and they appear either monolexematic or polylexematic structures. This paper shows that terminological reutilization doesn’t exist between these homophone countries in each terminographical instruments analyzed.

Key words: Archival science; Portuguese language, Provenance principle; Terminology.

‘Proveniencia’ en la terminografia archivística de lengua portuguesa: minería de textos y visualización de similitudes en términos y definiciones

Resumen: Este artículo se centra en el término ‘proveniência’ (procedencia), basado en la terminología archivística publicada en Portugal y Brasil (1986-2013). Se analiza la similitud/disimilitud entre las entradas terminológicas y las definiciones basadas en un *corpus* diacrónico. Se utilizó un enfoque semasiológico utilizando técnicas básicas de minería de texto y métodos de agrupamiento de texto con una visualización en árbol. El término 'proveniência' ocurre en 71 (a partir de 2760) entradas y aparecen estructuras monolexemáticas o polilexemáticas. Este trabajo muestra que no existe reutilización terminológica entre estos países homófonos en cada uno de los instrumentos terminológicos analizados.

Palabras-claves: Archivística; Ciencia de Archivos; Lengua portuguesa; Princípio de procedencia. Terminologia.

1 Introdução

A Arquivística tem vindo a afirmar-se como disciplina científica dos arquivos desde finais do século XIX (JENKINSON, 1922; MULLER; FEITH; FRUIN, 1898; THOMASSEN, 2015). Como tal, a arquivística ou a arquivologia – a primeira mais frequente em Portugal, a segunda mais frequente no Brasil, ainda que ambos os termos ocorram como equivalentes de acordo com determinada produção científica brasileira (BRITO, 2005; TOGNOLI; GUIMARÃES, 2011) – continua a ser objeto de intensa discussão, designadamente sob o conspecto do contexto pós-moderno e pós-custodial (FONSECA, 2008; SOARES; PINTO; SILVA, 2016). Se, para determinados autores, a arquivística é entendida como disciplina aplicada da Ciência da Informação (CI) (RIBEIRO, 2011), para outros constitui uma disciplina científica *ex proprio jure* (GILLILAND, 2017). O presente artigo, contudo, não se propõe analisar dicotomias entre as perspetivas custodialistas e pós-custodialistas, remetendo-se para os autores supramencionados. O que importa realçar é que a arquivística desenvolveu um vocabulário especializado que representa a teoria e a prática sobre os arquivos (BNP, 2010), partindo de um conceito fundacional como <proveniência>.

A terminologia arquivística, na perspetiva de Ribeiro (2001), está associada a um processo de estruturação epistemológica da disciplina baseada numa visão patrimonialista, historicista, custodial e tecnicista, centrada no universo analógico e documentrica, periodologicamente situada entre a publicação do manual holandês (MULLER; FEITH; FRUIN, 1898) e o ponto de viragem lançado por Ham (1981). Atualmente, assiste-se a um influxo terminológico muito significativo, devido não apenas ao incremento e diversificação dos novos *media* e ecossistemas digitais como também em concorrência com outras áreas epistémicas, como as Ciências da Computação, as Ciências Empresariais, as Ciências Jurídicas, a Filosofia, a Ciência Terminológica a própria Ciência da Informação (DÍEZ CARRERA, 2011; HEREDIA HERRERA, 2011). Esta inflação de termos obriga a uma constante revisão terminológica e tem consequências onto-epistemológicas para a arquivística, impelindo-a para o terreno da transdisciplinaridade (GILLILAND; MCKEMMISH; LAU, 2017; MCKEMMISH; GILLILAND, 2013).

O termo 'proveniência' está associado a um princípio fundamental da arquivística, onde gravitam vários conceitos, perspetivas e contradições (BARTLETT, 1992; GUIMARÃES; TOGNOLI, 2015; LEMIEUX, 2016). Baseia-se no pressuposto de um conjunto informacional ou fundo, produzido e/ou acumulado por uma entidade, não deve ser misturado com outro(s) fundo(s) (DUCHEIN, 1998; MICHETTI, 2016), conforme se pode ver no Quadro 1, na Legenda 1. Este princípio radica, por um lado, na tradição germânica do *Provenienzprinzip* (SPIESS, 1777) e, sobretudo, na tradição francesa de *respect des fonds* (DUCHÂTEL, 1841).

Quadro 1: Seleção de definições de 'proveniência' na terminografia arquivística internacional (inglês)

Entrada terminológica	Definição	Fonte
provenance	"1. The origin or source of something. - 2. Information regarding the origins, custody, and ownership of an item or collection. Notes: Provenance is a fundamental principle of archives, referring to the individual, family, or organization that created or received the items in a collection. The principle of provenance or the respect des fonds dictates that records of different origins (provenance) be kept separate to preserve their context."	(PEARCE-MOSES, 2005)
principle of provenance	"The basic principle that records/archives of the same provenance must not be intermingled with those of any other provenance; frequently referred to as "respect des fonds".	(ICA, 2004)
provenance	"IP2: The relationships between records and the organizations or individuals that created, accumulated and/or maintained and used them in the conduct of personal or corporate activity. [Archives]"	(DURANTI *ET AL* 2016)

Fontes: DURANTI *et al*., 2016; ICA, 2004; PEARCE-MOSES, 2005.

Na terminografia arquivística de língua portuguesa coexistem diferenças terminológicas entre Portugal e Brasil, devido a tradições sócio-culturais e jurídico-administrativas distintas (GUIMARÃES; TOGNOLI, 2015; SILVA *et al*., 2015). Apesar de os estudos sobre terminologia arquivística em português serem escassos, uns cingiram-se ao *corpus* terminográfico brasileiro (BALMANT, 2016; BELLOTTO, 2011; FROTA, 2015; RANGEL, 2015; ROCHA, 2011; SIQUEIRA, 2011), outros analisaram conceitos sob o macrodomínio da CI (MEDEIROS, 1986; UNIVERSIDADE DO PORTO, 2007) ou, ainda, centraram-se em problemas tradutológicos (SILVA *et al.*, 2015; TOGNOLI; GUIMARÃES; CÂNDIDO, 2016).

O problema de investigação deste estudo consiste em compreender as ocorrências de 'proveniência' na terminografia arquivística publicada em Portugal e no Brasil, entre 1986 a 2013. Como tal, a prospeção de dados não estruturados (*text mining*) constitui um dos métodos para processamento de dados textuais para suporte a uma análise semasiológica baseada em *corpora*. Este estudo tem como objetivo identificar e comparar as ocorrências do termo 'proveniência' em relação a similaridades e dissimilaridades entre textos através de um processo de aglomeração hierarquizada (*document clustering*). O objeto de análise consiste em termos e definições da terminografia arquivística para permitir um estudo comparativo das ocorrências do termo 'proveniência'.

2 Contextualização

2.1 Terminologia: conceitos, termos e definições

A terminologia, como disciplina científica, teve o seu próprio percurso teórico-praxeológico. Por um lado, a Teoria Geral da Terminologia (TGT) propôs uma perspetiva prescritiva e onomasiológica relativamente às conexões conceito-termo (WÜSTER, 1979; WÜSTER *et al*., 1979). Tais pressupostos teórico-metodológicos foram adotados por organismos internacionais de normalização (CAMPO, 2013). Por outro, a Teoria Comunicativa da Terminologia (TCT), distintamente da TGT, encetou uma abordagem de análise semasiológica das unidades terminológicas (CABRÉ, 1999, 2003). Outras pespetivas contribuíram para o desenvolvimento de novas linhas teórico-metodológicas como a socioterminologia (GAUDIN, 1990, 1993, 2005; TEMMERMAN, 2000), a terminologia baseada na semântica de *frames* (FABER, 2015), a teoria sociocognitiva da terminologia (TEMMERMAN, 2001), a ontoterminologia (ROCHE, 2012; ROCHE *et al*., 2009), entre outras abordagens sincréticas (COSTA, 2013; SANTOS; COSTA, 2015). Atualmente, a ciência terminológica constitui um campo interdisciplinar, recebendo contributos fundamentais sobretudo da linguística computacional (CABRÉ, 2003; IBEKWE-SANJUAN; CONDAMINES; CABRÉ, 2007) e da engenharia do conhecimento (LACASTA; NOGUERAS-ISO; ZARAZAGA-SORIA, 2010), constituindo-se como uma das áreas mais dinâmicas do *cluster* científico denominado Humanidades Digitais (EBENSGAARD JENSEN, 2014).

A terminologia, nas acepções dadas pela norma ISO 1087-1:2000, consiste (i) numa ciência que estuda a estrutura, a formação, o desenvolvimento, o uso e a gestão de termos utilizados em contextos discursivos específicos (ISO, 2000, pt. 3.5.2) e (ii) conjunto de designações pertencentes a uma linguagem especializada (ISO, 2000, pt. 3.5.1). As unidades fundamentais de análise em Terminologia são, portanto, termos, conceitos e definições (cf. *infra* Quadro 2). Por um lado, o conceito constitui uma unidade de representação do conhecimento, estabelecendo vínculos com os objetos e os signos de forma coerente. O conceito é o objeto principal na definição terminológica (DEPECKER, 2015). Por outro, os termos constituem denominações verbais dos conceitos e podem conter unidades superiores à palavra.

No caso da definição, a norma ISO 1087-1:2000 identifica duas estruturas tipológicas de definições: (i) definição intensional (ou analítica), que consiste na indicação da classe conceptual a que o termo pertence, na descrição de um conceito geral através das suas caraterísticas diferenciadoras e das suas relações de contiguidade com outros conceitos; e (ii) definição por extensão, por seu turno, respeita à descrição de conceitos por meio da enumeração de todos os conceitos subordinados, em função de um critério de subdivisão ou de uma descrição denotativa (NILSSON, 2015). Refira-se que as tipologias de definições radicam na tradição aristotélico-escolástica (em grego antigo, *ὁρισμός*) (ARISTÓTELES, 2007, pt. 101b39), que se estrutura numa sequência entre um termo (*ὅρος*/*definiendum*) e a sua definição (*ὁρισμός*/*definiens*). A definição, neste sentido, é construída com base na

delimitação dos seus elementos ou propriedades constitutivas, *i. e.*, *γένος/genus* e *διαφορά/differentiae specificae* (DESLAURIERS, 2007; LIDDELL; SCOTT, 1940a, 1940b).

Quadro 2: Seleção de entradas e definições metaterminológicas segundo iso 1087 (2000)

Termo	Definição	Fonte
concept	"unit of knowledge created by a unique combination of characteristics (…)"	ISO 1087-1:2000(en), 3.2.1
definition	"representation of a concept (…) by a descriptive statement which serves to differentiate it from related concepts"	ISO 1087-1:2000(en), 3.3.1
term	"verbal designation (…) of a general concept (…) in a specific subject field (…)"	ISO 1087-1:2000(en), 3.4.3

Fonte: INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2000.

No entanto, as perspetivas em torno dos conceitos, termos e definições não são consensuais (mesmo entre terminólogos), que variam em função de determinados campos científicos como a linguística, a lógica, a inteligência artificial ou a engenharia do conhecimento. Para melhor elucidação sobre a diversidade de posições, remetemos a título de complemento para Cabré (2003), Costa (2013), Kageura (2015), Santos e Costa (2015).

2.2 Terminologia no domínio da arquivística

A normalização do vocabulário arquivístico constituiu uma necessidade demonstrada desde finais do século XIX (JENKINSON, 1922; MULLER; FEITH; FRUIN, 1898). Após a institucionalização do Conselho Internacional dos Arquivos (1948, doravante ICA), encetaram-se esforços para a construção de uma terminologia multilíngue (ICA, 1964, 1984, 2004, 2013). No entanto, além das idiossincrasias terminológicas verificadas entre países heterófonos, verificou-se, também, que em contexto intralinguístico existiam diferenças devido a tradições jurídico-administrativas e culturais muito distintas entre esses países (DRYDEN, 2005).

A necessidade de harmonização da terminologia arquivística em língua portuguesa foi lançada por Mário Alberto Nunes Costa (1920-2010) em 1968, com uma proposta de criação de uma comissão de normalização terminológica Portugal-Brasil, para "coligir as noções arquivísticas em língua portuguesa, sejam tradicionais ou recentemente adquiridas, atingir definições, unificar, quando necessário, a terminologia" (NUNES, 1968, p. 7). No entanto, Portugal (ALVES; RAMOS; GARCIA, 1993; APBAD, 2001; BNP, 2010) e Brasil (ABNT, 1986; ARQUIVO NACIONAL, 2005; CAMARGO; BELLOTTO, 1996; DANNEMANN *et al.*, 1972; NAGEL, 1989; ROCHA, 2011) realizaram percursos distintos, ainda que o *DeltCI* do *Observatório de*

Ciência da Informação constitua uma exceção (UNIVERSIDADE DO PORTO, 2007). No caso dos países africanos de língua oficial portuguesa, Timor-Leste e Macau, não se conhecem *instrumenta* terminográficos publicados neste âmbito. Para suprir as diferenças terminológicas entre a Comunidade dos Países de Língua Portuguesa (CPLP), os acordos firmados no *Fórum dos Arquivos de Língua Portuguesa* (FALP), criado em 2003, incluíram na *Proposta de Resolução sobre a Cooperação na Área de Arquivos*, anexa à ata número 1, a alínea d): "contribuir para a fixação de um 'corpus' de terminologia arquivística em língua portuguesa" (CPLP. FALP, 2003). A não existência de *instrumenta* terminográficos sobre arquivística para o âmbito da CPLP impossibilita de algum modo a reutilização de termos e de definições de âmbito intralinguístico? É possível efetuar uma harmonização terminológica sem afetar o património terminológico de cada país?

O termo 'proveniência' ocorre na terminografia arquivística de língua portuguesa sob diversas formas, não só nos termos representados em entradas terminológicas como também nas definições, ora de forma monolexemática (u. g., 'proveniência') ora polilexemática (u. g., 'princípio da proveniência', 'proveniência territorial'), mantendo-se como núcleo o conceito <proveniência>. Em primeiro lugar, o termo em estudo é entendido pela comunidade arquivística como a materialização do fundo oriundo de um produtor, apesar de constituir um 'silo' informacional, representado como um sistema estático e unidimensional, verificável em vários instrumentos internacionais para a descrição arquivística e em normas de descrição emanadas por entidades responsáveis pela política arquivística nacional (BISWAS; SKENE, 2016; DAINES *et al.*, 2011; MACNEIL, 2012; OLIVER, 2010). Em segundo lugar, outros autores discordaram desta perspetiva, dado que o conceito <proveniência> está associado a uma rede complexa de relações entre agentes, objetos e funções, emergindo a pespetiva de 'múltiplas proveniências' (MILLAR, 2002; THIBODEAU, 2016).

Para Guimarães e Tognoli (2015) e Tognoli *et al*. (2016), o termo 'proveniência' deve ser estudado no conspecto da análise de domínio segundo critérios funcionais, ou seja, de análise *top-down*. Recentemente, o ICA apresentou uma perspetiva multidimensional e de base ontológica para a representação da informação arquivística, materializado no *draft* do *RiC: Records in Context* (GUEGUEN *et al*., 2013; ICA, 2016), apesar de não encontrar-se declarados quais os modelos de representação semântica que melhor exprimam computacionalmente a proveniência, *u. g. PROV-O* (LEBO *et al*., 2013), CRM_{dig} (BOUNTOURI; GERGATSOULIS, 2011; THEODORIDOU *et al.*, 2010) ou *OPM: Open Provenance Model* (MOREAU *et al*., 2010).

3 Materiais e metodologia

Este estudo estabelece a seguinte questão de investigação: como visualizar termos e definições onde o termo 'proveniência' ocorre, especialmente no que diz respeito à similaridade/dissimilaridade de entradas/definições na terminografia arquivística em língua portuguesa?

Existem diversas abordagens baseadas tanto nos métodos da Terminologia (cf. ISO 1087-1:2000), cujo objetivo consiste em resolver a ambiguidade entre conceitos e suas denominações, como no âmbito da Linguística de *Corpus*, que se baseia no processamento computacional de dados textuais (HEYLEN; HERTOG, 2015; KÜBLER; ZINSMEISTER, 2015). De acordo com Sager (1990), o trabalho terminológico sustenta-se atualmente em métodos computacionais para processamento de *corpora* textuais. Para Hacken (2010), o processamento baseado em *corpora* linguísticos constitui um exercício carateristicamente semasiológico, porque é a partir dos termos contidos em textos especializados que se pode obter informação sobre se as relações lexicais e semânticas refletem ou não a existência de vínculos concetuais (SANTOS; COSTA, 2015). Por exemplo, a extração automática de termos em *corpora* linguísticos constitui um dos métodos de processamento de linguagem natural aplicados no âmbito da Língua Portuguesa (GAUDIO, 2013; GAUDIO; BRANCO, 2007; WENDT, 2011).

Para o presente estudo, conforme já atrás exposto, adotamos uma abordagem empírica de prospeção de dados em *corpora* textuais para a identificação de ocorrências do termo 'proveniência' na terminografia arquivística publicada em Portugal e no Brasil entre 1986 a 2013 (cf. *infra* Quadro 3). Propõe-se efetuar uma aproximação preliminar à prospeção de dados textuais (*text mining*) e visualização de padrões de (dis)similaridade entre textos legíveis por máquina (*datasets*) onde ocorre o termo 'proveniência'. Estes padrões são obtidos estatisticamente a partir de técnicas de aprendizagem de máquina não supervisionada. Esta técnica permite agrupar dados a partir de um algoritmo que os agrega autonomamente em classes, para suporte a uma análise descritiva dos resultados (AGGARWAL, 2012; AGGARWAL; ZHAI, 2012).

Deste modo, é possível realizar um estudo baseado em:

(i) dados de *corpora* e processamento computacional,

(ii) seleção de um método de agrupamento por similaridade de fontes (*datasets*) contendo entradas terminológicas e respetivas definições, e

(iii) visualização dendrífica para suportar uma análise comparada dos resultados (GOMAA; FAHMY, 2013; HUANG, 2008; NENADIC; SPASIS; ANANIADOU, 2004; TAYLOR, 2013).

O processamento de *corpora* proposto neste estudo difere das técnicas mais utilizadas em Linguística de *Corpus* e na Terminologia, e, por conseguinte, não integrarão no âmbito

deste trabalho. O propósito deste estudo consiste somente na aplicação de técnicas básicas de *text mining* para a classificação de documentos (*document clustering*) como suporte para uma análise comparada de (dis)similaridades entre textos. Deste modo, optámos pelo seguinte fluxo de trabalho:

a) ***Constituição de* corpora *e pré-processamento.*** A terminografia arquivística em português é recente, com uma variação tipológica significativa, como dicionários, glossários, listas de vocabulários, bases de dados relacionais e *thesauri* (GÓMEZ DÍAZ, 2010). O presente *corpus* (Quadro 3) compõe-se de textos impressos, que foram digitalizados com ativação de OCR (*optical character recognition*) e armazenados em formato editável não proprietário (.txt). Inclui neste *corpus* bases de dados em linha, designadamente AC, BE e BD. A extensão cronológica dos textos é de 1986 a 2013, o que constitui um *corpus* de natureza diacrónica (HEYER; NIEKLER, 2016). Os termos e respetivas definições foram processadas em .csv (*datasets* acessíveis em MACEDO, 2017), mantendo a entrada terminológica (ET) no metadado de título precedido de código de fonte. O conteúdo de cada *dataset* (correspondente a uma ET) contém as definições da fonte mencionada no metadado de título. Refira-se que as ET contêm termos que designam uma definição. O *corpus* é composto por sete fontes textuais com num total de 2760 ET, segmentadas em .txt (*text file*). Esta segmentação é indispensável para a constituição não apenas de uma *bag of words* mas também para obter uma classificação hierarquizada dos documentos em função da sequência de palavras, que é o objetivo deste estudo. Não integra neste *corpus*, apesar da sua importância, textos terminográficos como Dannemann et al. (1972) e os mencionados na bibliografia em Arquivo Nacional (2005) por não estarem acessíveis na Internet. Não se incluiu também APBAD (2001), UNIVERSIDADE DO PORTO (s.d.) e os glossários integrantes das normas internacionais do ICA, editadas em Portugal e no Brasil. Excluímos deste *corpus* a tradução do inglês para português realizada pela equipa brasileira no *InterPARES Trust* (DURANTI *et al.*, 2016), por não ter devolvido qualquer entrada terminológica no dicionário eletrónico e por os critérios de tradução não serem explícitos. Excluímos as equivalências de termos em português em relação às línguas extrangeiras.

Quadro 3: *Corpus* de fontes terminográficas em língua portuguesa

Código *dataset*	Fontes	Formato original	ISO 3166-1 (país/língua de publicação)	ET (n=2760)
AA	(ALVES *et al.,* 1993)	Impresso	PT	563
AB	(BNP, 2010)	Nadodigital (.pdf)	PT	170
BA	(ABNT, 1986)	Impresso	BR	65
BB	(NAGEL, 1989)	Impresso	BR	537
BC	(CAMARGO & BELLOTTO, 1996)	Impresso	BR	594
BD	(BRASIL. ARQUIVO NACIONAL, 2005)	Nadodigital (.pdf)	BR	595
BE	(ICA, 2013)	Base de dados, *web*	BR*	236

Fonte: conforme mencionado no quadro.
Notas: *baseado no glossário do *InterPARES 3* (ROCHA, 2011).

b) ***Carregamento e prospeção de dados textuais.*** Para podermos processar dados textuais em ambiente controlado, selecionámos o *software open source* estatístico R (R CORE TEAM, 2013) e instalação do *package* "tm" (FEINERER; HORNIK, 2015) em

ambiente Windows da Microsoft Corporation. A escolha desta aplicação permite realizar "questões" através de funções e parâmetros do *package* "tm" e obter "respostas" automatizadas, sendo *outputs* obtidos automaticamente. A prospeção de dados textuais com base nesta ferramenta processa-se em conformidade com o Quadro 4.

Quadro 4: fluxo de trabalho com recurso ao *package* tm (FEINERER; HORNIK, 2015)

	Tarefa	Função
1	importação dos *datasets*	`> cname <- file.path("diretorio"; "nome do ficheiro")`
2	confirmação de *pathway*	`> dir(cname)`
3	quantificação de *datasets* importados	`> docs <- Corpus(DirSource(cname))`
4	visualização de conteúdo de datasets	`> inspect(docs)`
5	pré-processamento (por remoção/uniformização)	
	▪ pontuação	`> docs <- tm_map(docs, removePunctuation)`
	▪ números	`> docs <- tm_map(docs, removeNumbers)`
	▪ minúsculas	`> docs <- tm_map(docs, tolower)`
	▪ espaços vazios	`> docs <- tm_map(docs, stripWhitespace)`
6	Constituição de matriz dtm	`> dtm <- DocumentTermMatrix(docs)`
7	Matriz TF-IDF	`> dtm <-DocumentTermMatrix(docs,control= list(weighting = function(x) weightTfIdf(x, normalize = FALSE)))`
8	Hierarquização dendrífica [package *hclust*]	`> hc=hclust()` `> plot(hc)`

Fonte: dados da pesquisa, cf. Macedo (2017).

Refira-se que não se efetuou a remoção de *stopwords* nem se aplicou a lematização (*stemming*), por o objetivo deste artigo consistir numa identificação preliminar de (dis)similaridades entre *datasets*. Após o pré-processamento, é possível constituir uma matriz documento-termo (*dtm*). Entende-se por *dtm* uma matriz matemática que descreve a frequência dos termos que ocorrem numa coleção de documentos (*datasets*), sendo que as linhas correspondem a documentos na coleção e as colunas correspondem a termos. Apesar de existirem vários métodos para a obtenção de ponderações sobre a distribuição entre termos e documentos, optámos pela medida estatística *term frequency-inverse document frequency* (TF-IDF), para identificar o peso de uma palavra num documento em relação a um conjunto de documentos. Um termo mais frequente (TF) não quer dizer que seja mais significativo, pelo que o cálculo inverso (IDF) reduz o peso de um termo mais frequente contido num texto ou num conjunto de textos. Esta medida estatística constitui uma das muitas fórmulas utilizadas para a extração automática de termos e para a recuperação de informação (HEYLEN; HERTOG, 2015). A fórmula expressa-se como *p=tf(f)*df(n)* e pode ser obtido através do argumento *> dtm <-DocumentTermMatrix(docs,control= list(weighting = function(x) weightTfIdf(x, normalize = FALSE)))* (FEINERER; HORNIK, 2015, pp. 50–51). Com base neste cálculo, pode-se obter uma visualização de resultados, através do *package hclust* usando os argumentos *hc=hclust()* e *plot(hc)*. Em cada documento onde ocorre o termo 'proveniência', agrupar-se-á hierarquicamente por ramos, sendo que os ramos mais próximos

entre si corresponderão a maior similaridade entre *datasets*. Os ramos mais distantes correspondem a documentos com menor similaridade.

4 Resultados e discussão

Com base no fluxo de trabalho atrás mencionado, o universo de análise foi de 2760 *datasets*. Constituímos dois *subcorpora:* o *subcorpus* 1 corresponde às ocorrências do termo 'proveniência' tanto nas entradas terminológicas como nas definições e o *subcorpus* 2 corresponde apenas às ocorrências do termo em estudo nas entradas terminológicas. O Quadro 5 exemplifica, a título de síntese, dados obtidos a partir da constituição de *dtm*, a saber:

Quadro 5: Resultados da matriz documento-termo

Resultados	Ocorrências do termo 'proveniência'	
	Subcorpus 1 (ET + definições)	**Subcorpus 2 (Termos em ET)**
Documents	71	24
Terms	690	240
Non-/sparse entries	1721/47269	505/5255
Sparsity	96%	91%
Maximal term length	23	23

Fonte: dados da pesquisa, cf. Macedo (2017).

Se focalizarmos apenas para termos onde 'proveniência' ocorre no *subcorpus* 2, apenas 24 *datasets* apresentam este resultado, com um total de 240 unidades lexicais (*terms*). Verifica-se que 'proveniência' se apresenta quer de forma monolexemática – *u. g.,* 'proveniência'– quer polilexemática, como 'contexto de proveniência', 'princípio da proveniência', 'procedência/proveniência', 'proveniência funcional', 'proveniência territorial'.

Para identificar as associações de contiguidade deste termo, a partir das definições do subcorpus 2, com outras unidades lexicais, utilizámos o argumento *> findAssocs(dtm, "proveniência", corlimit=0.1)*, com os seguintes resultados:

a. ponderação igual a 0.57: 'proveniência' está associada a unidades lexicais que exprimem o resultado de processos de acumulação informacional (*conservados*, *derivado, elaborados, recebidos, resultantes*) e associado a um ciclo de vida (*corrente, intermédio, definitivo*).

b. ponderação inferior a 0.45: proveniência vinculada a relações multientidades como objetos (*arquivo* (0.22), *documentos* (0.12)), agentes (*produtora* (0.43), *entidade* (0.38)) e funções (*pertinência* (0.43), *funcional* (0.16), *territorial* (0.38)).

Por outro lado, as relações entre *procedência* e *proveniência* (BD, BB) manifestaram-se de forma ambígua (BB), cuja diferenciação foi esclarecida posteriormente em BD. Este facto só é verificável com recurso a *corpora* diacrónicos do mesmo domínio para contextualização

evolutiva dos conceitos (HEYER *et al.*, 2016), uma vez que o papel da terminologia consiste em resolver ambiguidades. No caso de entradas *proveniência territorial* e *proveniência funcional*, apresentam-se como hipónimos de 'proveniência'. Noutros casos, 'proveniência' apresenta-se como sinónimo de *pertinência* e *procedência*, ainda que 'procedência' apresente diferenças semânticas em relação ao termo 'proveniência' (BD), conforme Quadro 6.

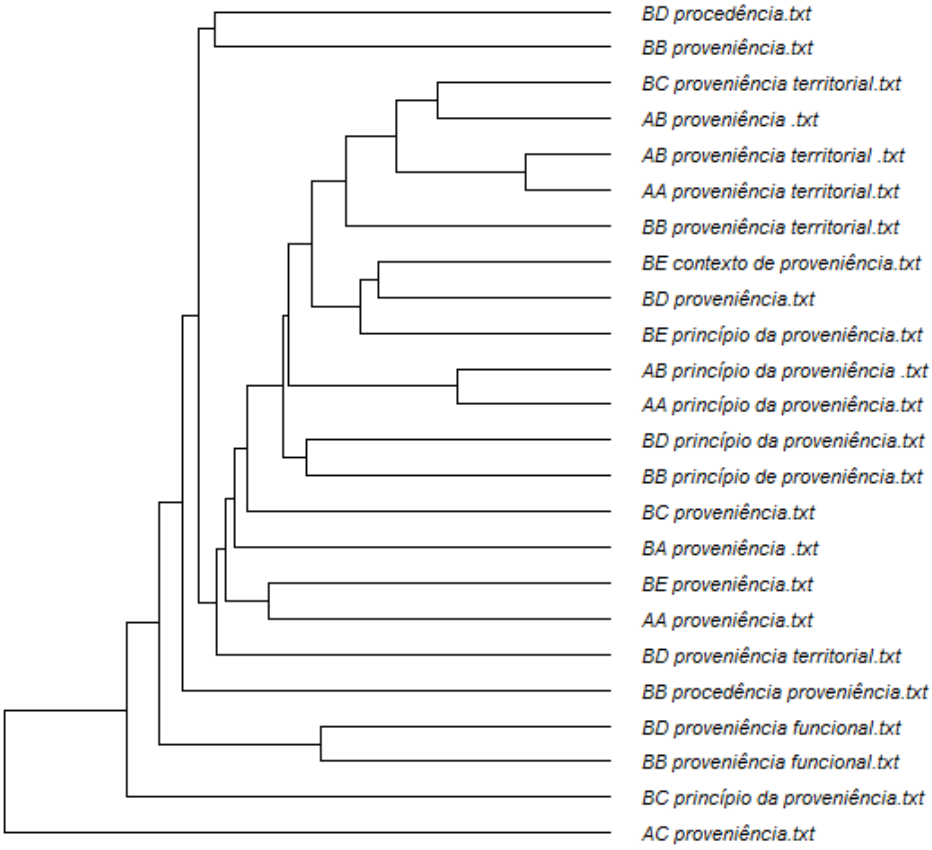
Quadro 6: Seleção de entradas terminológicas (BR)

Fonte	Código	Definição
BB	Procedência / proveniência	"Instituição, administração, estabelecimento, organismo ou pessoa privada que criou, acumulou e conservou documentos de arquivo durante a realização de seus negócios antes de sua transferência a um centro de pré-arquivo ou a um Arquivo."
BB	proveniência	"Origem, procedência de papéis de arquivo que no curso das atividades foram acumulados, preservados por um órgão, administração ou pessoa privada que os criou no curso de suas ações."
BD	procedência	"Termo em geral empregado para designar a origem mais imediata do arquivo, quando se trata de entrada de documentos efetuada por entidade diversa daquela que o gerou. Conceito distinto do de proveniência."
BD	Proveniência	"Termo que serve para indicar a entidade coletiva, pessoa ou família produtora de arquivo."

Fonte: dados da pesquisa, cf. Macedo (2017).

Para obter relações de similaridade/dissimilaridade entre *datasets* de ET (*subcorpus* 2), a partir dos resultados do cálculo TF-IDF da matriz *dtm*, é possível obter uma visualização dendrífica com recurso ao package *hclust* e do argumento > *hc=hclust(dist(dtm)).* Este tipo de visualização por aglomerações (*clusters*) permite-nos testar hipóteses de (dis)similaridade entre textos, o que facilita direcionar a atenção para *clusters* mais significativos, *i. e.*, potenciar uma análise comparada de ET que apresentam da maior à menor similaridade nas suas definições. Tal pode ser observado no dendrograma da Figura 1.

Figura 1: Dendrograma de fontes terminográficas com ocorrência mono e polilexemática do termo 'proveniência' do *subcorpus* 2.



Fonte: dados da pesquisa, cf. Macedo (2017).

O mais significativo, a nosso ver e a título exemplificativo, consiste nos pares com maior similaridade, sobretudo os que contêm a mesma sequência de *tokens*, como *proveniência territorial* e *princípio da proveniência* (AB, AA) e *proveniência funcional* (BD, BB), como se pode aferir na Quadro 7.

Neste caso, assiste-se a uma reutilização de termos e de definições dentro da tradição terminológica de cada país, exceto *proveniência* (BD, AB). A terminografia portuguesa (A) e brasileira (B) aparece, na sua maioria em *clusters* distintos (cf. supra Figura 1).

Quadro 7: Extrato exemplificativo de similaridade entre *datasets*

Fonte	ET	Definição
AA	*proveniência territorial*	*Conceito segundo o qual os arquivos {1} devem ser mantidos sob a jurisdição arquivística do território onde foram produzidos.*
AB	*proveniência territorial*	*Conceito segundo o qual os arquivos devem ser mantidos sob a jurisdição arquivística do território onde foram produzidos.*
AA	*princípio da proveniência*	*Princípio básico da organização de arquivos {1}, segundo o qual deve ser respeitada a autonomia de cada fundo ou núcleo, não misturando os seus documentos com os de outros.*
AB	*princípio da proveniência*	*Princípio básico da organização, segundo o qual deve ser respeitada a autonomia de cada arquivo, não misturando os seus documentos com os de outros.*
BD	*proveniência funcional*	*Conceito segundo o qual, com a transferência de funções de uma autoridade para outra como resultado de mudança política ou administrativa, documentos relevantes ou cópias são também transferidos para assegurar a continuidade administrativa. Também chamado pertinência funcional.*
BB	*proveniência funcional*	*Conceito segundo o qual, quando há transferência de funções de uma autoridade a outra como resultado de mudança políticas ou administrativas, os documentos relevantes ou cópias deles devem ser também transferidos, a fim de garantir a continuidade administrativa. Também referido como pertinência funcional.*

Fonte: dados da pesquisa, cf. Macedo (2017).

Da leitura feita em torno dos resultados da prospeção de dados textuais com enfoque no conceito arquivístico de <proveniência>, os dados apontam para o facto de, em contexto intralinguístico, Portugal e Brasil corresponderem a duas tradições distintas de produção terminográfica. Tal divergência poderá estar associada às políticas de normalização que se cingem ao perímetro de cada nação e às políticas linguísticas, com escasso ou nulo envolvimento das comunidades lusófonas e de outras áreas epistémicas. Além disso, a produção terminográfica de ambos os países tem adotado uma prática mais prescritiva do que descritiva, realizada de forma pontual no tempo e no espaço e não de forma sistemática e dinâmica, fundamentada na literatura científica e jurídico-normativa. Por exemplo, a definição dada ao termo 'proveniência territorial' (cf. supra Quadro 7) dá ênfase a um sistema arquivístico baseado numa proveniência que se estrutura por jurisdições territoriais. Contudo, as definições dadas a este termo não clarificam a sua relação em casos de arquivos expatriados (internacional e intranacionalmente), quer aplicáveis a bens culturais documentais quer ao armazenamento e acesso ubíquo de dados e informações na nuvem (vulgo, *cloud*).

Conforme exposto na contextualização do tema mais acima, o conceito de <proveniência> tem evoluído consideravelmente a partir da perspetiva da arquivística pós-custodial, especialmente nas *múltiplas proveniências*. Contudo, tais perspetivas não têm sido incorporadas nos mais recentes *instrumenta* terminográficos de língua portuguesa, conforme

se pode aferir a partir das definições atrás exaradas.

5 Conclusões

Este artigo efetuou uma análise contextual sobre terminologia e terminografia arquivística e, de forma exploratória, utilizámos técnicas básicas de prospeção de dados textuais em *corpora* de textos provenientes da terminografia arquivística, tendo como referência o termo 'proveniência'.

Foi possível verificar que as técnicas aplicadas neste estudo constitui uma hipótese de trabalho de identificação de (dis)similaridades entre textos a partir das definições. Da comparação efetuada em torno de (dis)similaridade(s) entre *datasets*, há aspetos positivos mas também contámos com limitações, designadamente:

1. a prospeção de dados textuais a partir de métodos de aprendizagem não-supervisionados permite-nos realizar um processamento de *corpus* em ambiente computacionalmente controlado. Apesar de estas técnicas serem emergentes no âmbito da Linguística de *Corpus*, exige um lato domínio de ferramentas de processamento de linguagem natural e bem como dos algoritmos de agrupamento (*clustering*);
2. embora o recurso a um método estatístico (TF-IDF) nos permita vislumbrar potencialidades de análise baseada em *corpora* textuais e hierarquização de fontes baseada no critério de similaridade, a diversidade de algoritmos de aglomeração hierarquizada poderá trazer resultados distintos, em função dos objetivos de prospeção (HUANG, 2008);
3. a análise a partir de *corpora* diacrónicos, como o caso observado com o termo 'proveniência', possibilita realizar um estudo em torno da evolução das relações semânticas e lexicais de forma incremental e combinada com outros termos;
4. apesar de este estudo ter-se centrado apenas em instrumentos terminográficos sobre arquivística publicados em português, não deixa de ser indispensável alargar, em trabalhos futuros, a pesquisa para *corpora* textuais de natureza jurídico-normativa e
5. académica.

A terminografia arquivística em língua portuguesa, conforme exposto, tem sido concebida numa ótica predominantemente prescritiva no seio da comunidade arquivística de ambos os países. Os resultados atrás expostos podem auxiliar a comparação e avaliação de instrumentos terminográficos a partir de métodos de técnicas de aglomeração de documentos (*document/text clustering*), que podem servir, a título de exemplo, de suporte às orientações da norma ISO 23185 (ISO, 2009).

Torna-se indispensável, em última análise e para concluir, que o estudo terminológico no domínio arquivístico deve procurar sustentar-se interdisciplinarmente com os métodos da Terminologia, da Linguística de *Corpus*, da Engenharia do Conhecimento e das Humanidades Digitais. Perante a rápida evolução, diversidade e quantidade de ecossistemas informacionais, a terminologia arquivística, designadamente em contexto intralinguístico, requer uma

abordagem descritiva, suportada com dados e, preferencialmente, com ampla participação no quadro da CPLP.

Agradecimentos

Agradecemos à Professora Doutora Rute Costa (Universidade Nova de Lisboa), à Professora Doutora Maria Manuel Borges (Universidade de Coimbra) e à Professora Doutora Maria Cristina Vieira de Freitas (Universidade de Coimbra) e aos revisores anónimos pelas notas e propostas sobre o presente texto.

Referências

AGGARWAL, C. C.; ZHAI, C. **Mining Text Data**. New York: Springer, 2012. Disponível em: https://doi.org/10.1007/978-1-4614-3223-4 Acesso em: 2017-04-04.

ALVES, I.; RAMOS, M. M. O.; GARCIA, M. M. **Dicionário de terminologia arquivística**. Lisboa : Instituto da Biblioteca Nacional e do Livro, Organismo de Normalização Sectorial para a Informação e Documentação, 1993.

ASSOCIAÇÃO PORTUGUESA DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS (APBAD). Terminologia Audiovisual. **Cadernos BAD**, v. 1, p. 1–98, 2001. Disponível em: http://www.bad.pt/publicacoes/index.php/cadernos/article/view/885/883 Acesso em: 2017-04-04.

ARISTÓTELES. ***Tópicos***. [Em linha]. Lisboa: Imprensa Nacional, Casa da Moeda, 2007. Disponível em: http://www.obrasdearistoteles.net/ Acesso em: 2017-04-04.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 9578: Arquivo: terminologia**. Rio de Janeiro: ABNT, 1986.

BARTLETT, Nancy. Respect des fonds: the origins of the modern archival principle of provenance. **Primary Sources & Original Works**, v. 1, n. 1–2, p. 107–115, 1992. Disponível em: http://dx.doi.org/10.1300/J269V01N01_07 Acesso em: 2017-04-04.

BALMANT, F. V. **Terminologia arquivística brasileira estudo exploratório de publicações e termos**. 2016. 252p. Tese (Mestrado) - Universidade Federal do Rio de Janeiro, Rio de Janeiro. Disponível em: http://www.unirio.br/ppgarq/tccs/turma-2013/BALMANT-%20Fabricio%20Vieira.%20Terminologia%20Arquivistica%20%20Brasileira..pdf/at_download/file Acesso em: 2017-11-29.

BELLOTTO, Heloisa Liberalli. A Terminologia das Áreas do Saber e do Fazer: O caso da arquivística. **Revista Acervo**, v. 20, n. 1/2, p. 47–56, 2011.

BIBLIOTECA NACIONAL DE PORTUGAL. **Normas portuguesas de documentação e informação CT 7**. Lisboa: BNP, IPQ, 2010.

BISWAS, Paromita; SKENE, Elizabeth. From Silos to (Archives)Space: Moving Legacy Finding Aids Online as a Multi-Department Library Collaboration. **The Reading Room: A Journal of Special Collections**, v. 1, n. 2, p. 67–84, 2016. Disponível em: http://readingroom.lib.buffalo.edu/readingroom/PDF/vol1-issue2/reading-room-vol1-issue2.pdf Acesso em: 2017-04-04.

BOUNTOURI, Lina; GERGATSOULIS, Manolis. The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology. **Journal of Archival Organization**, v. 9, n. 3–4, p. 174–207, 2011. Disponível em: DOI: 10.1080/15332748.2011.650124 Acesso em: 2017-04-04.

BRASIL. ARQUIVO NACIONAL. **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro : Arquivo Nacional, 2005. Disponível em: http://www.conarq.arquivonacional.gov.br/images/publicacoes_textos/dicionrio_de_terminologia_arquivistica.pdf Acesso em: 2017-04-04.

CABRÉ, María Teresa. **Terminology: Theory, methods and applications**. Amsterdam/Philadelphia: John Benjamins Publishing, 1999.

CABRÉ, María Teresa. Theories of terminology: their description, prescription and explanation. **Terminology**, v. 92, p. 163–199, 2003. Disponível em: https://doi.org/10.1075/term.9.2.03cab Acesso em: 2017-04-04.

CAMARGO, Maria Albertina; BELLOTTO, Heloísa Liberalli (Ed.). **Dicionário de terminologia arquivística**. São Paulo: Associação dos Arquivistas Brasileiros, Núcleo Regional de São Paulo; Secretaria de Estado da Cultura, 1996.

CAMPO, Angela. **The Reception of Eugen Wüster's Work and the Development of Terminology**. 2013. 378p. Tese (Doutorado) - Université de Montréal, Montréal. Disponível em: http://hdl.handle.net/1866/9198 Acesso em: 2017-04-04.

COSTA, Rute. Terminology and Specialized Lexicography: two complementary domains. **Lexicographica: Revue Internationale de Lexicographie**, v. 29, n. 1, p. 29–42, 2015. doi: 10.1515/lexi-2013-0004. Acesso em: 2017-04-04.

CPLP. FALP. **Proposta de Resolução sobre a Cooperação na Área de Arquivos: ata n.o 1** (2003-10-08 e 09). Lisboa: ANTT, 2003.

DAINES III, J. Gordon; NIMER, Cory L. Re-Imagining Archival Display: Creating User-Friendly Finding Aids. **Journal of Archival Organization**, v. 9, p. 4–31, 2011. doi: 10.1080/15332748.2011.574019 Acesso em: 2017-04-04.

DANNEMANN, M. L. S. *et al*. Terminologia arquivística. In: CONGRESSO DE ARQUIVOLOGIA, 1, Rio de Janeiro, 1972. Rio de Janeiro: Associação de Arquivistas Brasileiros, 1972. p. 435–495.

DEPECKER, L. How to build terminology science? In: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing, 2015. p. 34–44.

DESLAURIERS, Marguerite. **Aristotle on definition**. Leiden; Boston: Brill, 2007.

DÍEZ CARRERA, Carmen. Estudio terminológico y metodología aplicada. In: CRUZ MUNDET, R. (Ed.). **Diccionario de Archivística**. Madrid: Alianza Editorial, 2011. p. 13–46.

DRYDEN, Jean. A tower of Babel: standardizing archival terminology. **Archival Science**, v. 5, n. 1, p. 1–16, 2005. doi: 10.1007/s10502-005-9001-3 Acesso em: 2017-04-04.

DUCHÂTEL, T. **Rapport au roi sur les archives départementales et communales**. Paris: Imprimerie Royale, 1841.

DUCHEIN, Michel. Le principe de provenance et la pratique du tri, du classement et de la description en archivistique contemporaine. **Ligall: revista catalana d'Arxivística**, v. 1, n. 12, p. 87-100, 1998.

DURANTI, L. *et al*. **InterPARES Trust Terminology Database**. Disponível em: http://arstweb.clayton.edu/interlex/pt/ Acesso em: 10 nov. 2017.

EBENSGAARD JENSEN, Kim. Linguistics and the digital humanities: (computational) corpus linguistics. **Mediekultur: Journal of Media & Communication Research**, v. 30, n. 57, p. 115–134, 2014. Disponível em: http://dx.doi.org/10.7146/mediekultur.v30i57.15968 Acesso em: 2017-04-04.

FABER, Pamela. Frames as a framework for terminology. In: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing, 2015. p. 14–33.

FEINERER, Ingo; HORNIK, Kurt. **tm: Text mining package**. R package version 0.5-7.1. 2012;1(8).

FROTA, Bianca Celistre. **Análise comparativa de termos arquivísticos em língua portuguesa**. 2015. 53 p. Trabalho de fim de curso - Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: http://hdl.handle.net/10183/135060 Acesso em: 2017-04-04.

GAUDIN, François. La socioterminologie. **Langages**, v. 1, p. 81–93, 2005. doi: 10.3917/lang.157.0081 Acesso em: 2017-04-04.

GAUDIN, François. **Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles**. Rouen: Université de Rouen, 1993. Disponível em: http://hal-01090348 Acesso em: 2017-04-04.

GAUDIN, François. Socioterminology and expert discourses. **TKE'90: Terminology and knowledge engineering**, v. 2, p. 631–641, 1990. Acessível em: http://hal-01090697 Acesso em: 2017-04-04.

GAUDIO, R. Del; BRANCO, A. Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach. In: NEVES J.; SANTOS, M.F.; MACHADO, J.M. (eds). **Progress in Artificial Intelligence EPIA 2007**. Berlin, Heidelberg: Springer, 2007. p. 659–670 (Lecture Notes in Computer Science, v. 4874). DOI: 10.1007/978-3-540-77002-2_55 Acesso em: 2017-04-04.

GAUDIO, Rosa Del. **Automatic Extraction of Definitions**. 2013. 205p. Tese (Doutorado) - Universidade de Lisboa. Disponível em: http://hdl.handle.net/10451/10818 Acesso em: 2017-04-04.

GOMAA, Wael H.; FAHMY, Aly A. A survey of text similarity approaches. **International Journal of Computer Applications**, v. 68, n. 13, p. 13–18, 2013.

GÓMEZ DÍAZ, Raquel. Evaluación de herramientas terminológicas especializadas: El caso de los glosarios, diccionarios, índices analíticos y tesauros de archivística. In: GARCÍA PALACIOS, J. (Ed.). **La terminología de la archivística**. Gijón: Trea, 2010. p. 41–69.

GUEGUEN, Gretchen; FONSECA, Vitor da; PITTI, Daniel; GRIMOÜARD, Claire. Toward an International Conceptual Model for Archival Description: A Preliminary Report from the

International Council on Archives' Experts Group on Archival Description. **American Archivist**, v. 76, n. 2, p. 566–583, 2013. Disponível em: http://dx.doi.org/10.17723/aarc.76.2.p071x02401282qx2 Acesso em: 2017-04-04.

GUIMARÃES, José Augusto Chaves; TOGNOLI, Natália Bolfarini. Provenance as a Domain Analysis Approach in Archival Knowledge Organization. **Knowledge Organization**, v. 42, n. 8, p. 562–569, 2015.

HACKEN, Pius T. The Tension between Definition and Reality in Terminology. In: DYKSTRA, Anne; SCHOONHEIM, Tanneke (Ed.). **Proceedings of the XIV Euralex International Congress**. Ljouwert: Fryske Akademy/Afuk, 2010. p. 915–927. Disponível em: http://euralex.org/publications/the-tension-between-definition-and-reality-in-terminology Acesso em: 2017-04-04.

HAM, F. Archival strategies for the post-custodial era. **American Archivist**, v. 44, n. 3, p. 207–216, 1981.

HEREDIA HERRERA, Antonia. **Lenguaje y vocabulario archivísticos: algo más que un diccionario**. Sevilla: Junta de Andalucía, 2011.

HEYER, Gerhard; KANTNER, Cathleen; NIEKLER, Andreas; OVERBECK, Max; WIEDEMANN, Gregor. Modeling the dynamics of domain specific terminology in diachronic corpora. In: INTERNATIONAL CONFERENCE ON TERMINOLOGY AND KNOWLEDGE ENGINEERING (TKE 2016), AT COPENHAGEN BUSINESS SCHOOL, 12. Copenhagen: [s.n.]: 2016. p. 75–90. Disponível em: http://openarchive.cbs.dk/handle/10398/9323 Acesso em: 2017-04-04.

HEYLEN, Kris; HERTOG, Dirk. Automatic term extraction. In: KOCKAERT, H. J.; STEURS, F. (Eds.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015. p. 199-219.

HUANG, Anna. Similarity Measures for Text Document Clustering. In: NEW ZEALAND COMPUTER SCIENCE RESEARCH STUDENT CONFERENCE (NZCSRSC2008), 6th. **Proceedings**. Christchurch, New Zealand: 2008. p. 49-56.

IBEKWE-SANJUAN, Fidelia; Condamines, Anne; Cabré, Maria Teresa. **Application-driven Terminology Engineering**. Amsterdam/Philadelphia: John Benjamins Publishing, 2007.

INTERNATIONAL COUNCIL ON ARCHIVES. EGAD. **RiC-CM-0.1: Records in Contexts: a conceptual model for archival description**. Paris: ICA, 2016. Disponível em: http://www.ica.org/sites/default/files/RiC-CM-0.1.pdf Acesso em: 2017-04-04.

INTERNATIONAL COUNCIL ON ARCHIVES. **DAT III: Dictionary of archival terminology**. Disponível em: http://internet.archivschule.uni-marburg.de/datiii/index.html (última edição: 28. November 2004). Acesso em: 2017-04-04.

INTERNATIONAL COUNCIL ON ARCHIVES. **Dictionary of archival terminology: English and French; with equivalents in Dutch, German, Italian, Russian, and Spanish**. München: New York: K. G. Saur, 1984.

INTERNATIONAL COUNCIL ON ARCHIVES. **Elsevier's lexicon of archive terminology: French, English, German, Spanish, Italian, Dutch**. Amsterdam: Elsevier, 1964.

INTERNATIONAL COUNCIL ON ARCHIVES. **Multilingual Archival Terminology**. Disponível em: http://www.ciscra.org/mat Acesso em: 2017-04-04.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 1087-1:2000 Terminology work—Vocabulary. Part 1: Theory and application**. Geneva: International Organization for Standardization, 2000.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 23185: 2009 - Critères d'évaluation comparative des ressources terminologiques — Concepts, principes et exigences d'ordre général**. Geneva: International Organization for Standardization, 2009.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/TR 22134:2007 - Practical guidelines for socioterminology**. Geneva: International Organization for Standardization, 2007.

JENKINSON, Hilary. **A manual of archive administration including the problems of war archives and archive making**. Oxford: Clarendon Press, 1922.

KAGEURA, Kyo. Terminology and lexicography. In: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing, 2015. p. 45–59.

KÜBLER, Sandra; ZINSMEISTER, Heike. **Corpus Linguistics and Linguistically Annotated Corpora**. London: Bloomsbury Academic, 2015.

LACASTA, Javier; NOGUERAS-ISO, Javier; SORIA, Francisco Javier Zarazaga. **Terminological Ontologies: Design, Management and Practical Applications**. New York: Springer, 2010.

LEBO, Timothy; SAHOO, Satya; MCGUINNESS, Deborah; BELHAJJAME, Khalid; CHENEY, James; CORSAR, David; GARIJO, Daniel; SOILAND-REYES, Stian; ZEDNIK, Stephan; ZHAO, Jun. PROV-O: The PROV Ontology. **W3C recommendation**, v. 30, 2013. Disponível em: https://www.w3.org/TR/prov-o/ Acesso em: 2017-04-04.

LEMIEUX, Victoria L. (Ed.). **Building Trust in Information Perspectives on the Frontiers of Provenance**. Cham: Springer, 2016. DOI: 10.1007/978-3-319-40226-0 Acesso em: 2017-04-04.

LIDDELL, H. G.; SCOTT, R. - «διαφορ-ά , ἡ, (διαφέρω)». In: Jones, Henry Stuart. **A Greek-English Lexicon**. [S.l.]: Clarendon Press, 1940b. Disponível em: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Ddiafora%2F Acesso em: 2017-04-04.

LIDDELL, H. G.; SCOTT, R. «γένος , εος or ους, τό,». In: Jones, Henry Stuart. **A Greek-English Lexicon**. [S.l.]: Clarendon Press, 1940a. Disponível em: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Dge%2Fnos Acesso em: 2017-04-04.

MACEDO, L. S. Ascensão de. **Terminologia arquivística de língua portuguesa: *datasets* sobre «proveniência»**. [s. l.]: Harvard Dataverse, v. 1, 2017. Disponível em: https://doi.org/10.7910/DVN/UANDAH Acesso em: 2017-04-04.

MACNEIL, Heather. What finding aids do: Archival description as rhetorical genre in traditional and web-based environments. **Archival Science**. v. 12, n. 4, p. 485–500, 2012. DOI: 10.1007/s10502-012-9175-4 Acesso em: 2017-04-04.

MCKEMMISH, Sue; GILLILAND, Anne. Archival and recordkeeping research: Past, present and future. In: JOHANSON, Graeme; WILLIAMSON, Kirsty (Ed.). **Research Methods: Information, Systems, and Contexts**. Prahran, Vic: Tilde University Press, 2013. p. 79–112.

MEDEIROS, Marisa Brascher Basílio. Terminologia brasileira em Ciência da Informação: uma análise. **Ciência da Informação**, v. 15, n. 2, p. 135–142, 1986.

MICHETTI, Giovanni. Provenance: An Archival Perspective. In: LEMIEUX, Victoria L. (Ed.). **Building Trust in Information Perspectives on the Frontiers of Provenance**. Cham: Springer, 2016. p. 59–68.

MILLAR, Laura. The death of the fonds and the resurrection of provenance: archival context in space and time. **Archivaria**, v. 53, p. 1–15, 2002.

MOREAU, Luc; CLIFFORD, Ben; FREIRE, Juliana; FUTRELLE, Joe; GIL, Yolanda; GROTH, Paul; KWASNIKOWSKA, Natalia et al. The open provenance model core specification (v1.1). **Future Generation Computer Systems**, v. 27, n. 6, p. 743–756, 2011. Disponível em: http://dx.doi.org/10.1016/j.future.2010.07.005 Acesso em: 2017-04-04.

MULLER, S.; FEITH, J. A.; FRUIN, R. **Handleiding voor het ordenen en beschrijven van archieven: ontworpen in opdracht van de Vereeniging van Archivarissen in Nederland**. Groningen: Erven B. van der Kamp, 1898.

NAGEL, R. (Ed.). **Dicionário de termos arquivísticos: subsídios para uma terminologia arquivística brasileira**. Bonn: Salvador: Deutsche Stiftung für internationale Entwicklung: Universidade Federal da Bahia, 1989.

NENADIC, Goran; SPASIC, Irena; ANANIADOU, Sophia. Mining term similarities from corpora. **Terminology**, v. 10, n. 1, p. 55–80, 2004. doi: 10.1075/term.10.1.04nen Acesso em: 2017-04-04.

NILSSON, Henrik. Enumerations count: Extensional and partitive definitions. In: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing, 2015. p. 82 – 100.

NUNES, Mário Alberto Nunes. Apontamento para a unificação da terminologia arquivística na língua portuguesa. In: COLÓQUIO INTERNACIONAL DE ESTUDOS LUSO-BRASILEIROS, 5. Coimbra: Universidade de Coimbra, 1968. p. 5–8.

OLIVER, Gillian. Transcending silos, developing synergies: libraries and archives. **Information Research**, v. 15, n. 4, colis71, 2010. Disponível em: https://eric.ed.gov/?id=EJ912743 Acesso em: 2017-04-04.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Viena (Áustria): R Foundation for Statistical Computing, 2015. Disponível em: http://www.r-project.org Acesso em: 2017-04-04.

RANGEL, Kíssila da Silva. **Revisitando o princípio da proveniência: percepções sobre a organicidade**. 2015. 102p. Tese (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro. Disponível em: http://www.unirio.br/ppgarq/tccs/turma-2013/rangel-kissila-

da-silva-revisitando-o-principio-da-proveniencia-percepcoes-sobre-a-organicidade/at_download/file Acesso em: 2017-04-04.

RIBEIRO, Fernanda. A arquivística como disciplina aplicada no Campo da ciência da informação. **Perspectivas em Gestão & Conhecimento**, v. 1, n. 1, p. 59-73, 2011.

RIBEIRO, Fernanda. Archival science and changes in the paradigm. **Archival Science**, v. 1, n. 3, p. 295–310, 2001. Disponível em: DOI: 10.1007/BF02437693 Acesso em: 2017-04-04.

ROCHA, Cláudia Lacombe. Glossário multilíngue do Projeto InterPARES 3. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, p. 76-90, 2011. Disponível em: DOI: 10.5007/1518-2924.2011v16nesp1p76 Acesso em: 2017-04-04.

ROCHE, Christophe. Should Terminology Principles be re-examined? **arXiv preprint**, 2016. Disponível em: https://arxiv.org/abs/1609.05170v1 Acesso em: 2017-04-04.

ROCHE, Christophe; CALBERG-CHALLOT, Marie; DAMAS, Luc; ROUARD, Philippe. Ontoterminology: A new paradigm for terminology. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND ONTOLOGY DEVELOPMENT, 2009. p. 321-326. Disponível em: https://hal.archives-ouvertes.fr/hal-00622132 Acesso em: 2017-04-04.

SAGER, Juan C. **A practical course in terminology processing**. Amsterdam/Philadelphia: John Benjamins Publishing, 1990.

SANTOS, Claudia; COSTA, Rute. Domain specificity: Semasiological and onomasiological knowledge representation. In: KOCKAERT, H. J.; STEURS, F. (Ed.). **Handbook of Terminology**. Amsterdam/Philadelphia: John Benjamins Publishing, 2015. p. 153–179.

SILVA, Andrieli Pachu da; MOREIRA, Walter; GUIMARÃES, José Augusto Chaves; MORAES, João Batista Ernesto de. Organização do conhecimento arquivístico: um estudo terminológico comparativo (português, espanhol, francês, inglês) sobre classificação e descrição no Multilingual Archival Terminology. In: CONGRESO ISKO ESPAÑA Y II CONGRESO ISKO ESPAÑA-PORTUGAL, 12., 19-20 DE NOVIEMBRE, 2015, ORGANIZACIÓN DEL CONOCIMIENTO PARA SISTEMAS DE INFORMACIÓN ABIERTOS. Murcia: Universidad de Murcia, 2015. p. 1-9. Disponível em: http://www.iskoiberico.org/wp-content/uploads/2015/11/41_Silva.pdf Acesso em: 2017-04-04.

SIQUEIRA, Jéssica Camara. **As noções de documento e de informação - uma abordagem terminológica**. 2011. Dissertação (Mestrado em Cultura e Informação) - Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2011. DOI:10.11606/D.27.2011.tde-15122011-235031 Acesso em: 2017-04-04.

SPIESS, Philipp Ernst. **Von Archiven**. Hulle: bey Johann Jacob Gebauer, 1777.

TAYLOR, Charlotte. Searching for similarity using corpus-assisted discourse studies. **Corpora**, v. 8, n. 1, p. 81–113, 2013. DOI: 10.3366/cor.2013.0035 Acesso em: 2017-04-04.

TEMMERMAN, Rita. Sociocognitive Terminology Theory. In: CASTELLVÍ, María Teresa Cabré; FELIU, J. (Ed.). **Terminología y Cognición**. Barcelona: University Pompeu Fabra, 2001. p. 75–92.

TEMMERMAN, Rita. **Towards New Ways of Terminology Description. The Sociocognitive Approach**. Amsterdam/Philadelphia: John Benjamins Publishing, 2000.

THEODORIDOU, Maria; TZITZIKAS, Yannis; DOERR, Martin; MARKETAKIS, Yannis; MELESSANAKIS, Valantis. Modeling and querying provenance by extending CIDOC CRM. **Distributed and Parallel Databases**, v. 27, n. 2, p. 169–210, 2010. doi: 10.1007/s10619-009-7059-2 Acesso em: 2017-04-04.

THIBODEAU, Ken. Research Issues in Archival Provenance. In: LEMIEUX, Victoria L. (Ed.). **Building Trust in Information Perspectives on the Frontiers of Provenance**. Cham: Springer, 2016. p. 69–78.

THOMASSEN, Theo. Archival science. In: DURANTI, Luciana; FRANKS, Patricia C. (eds.). **Encyclopedia of archival science**. Lanham, MD: Rowman & Littlefield, 2015.

TOGNOLI, Natalia; GUIMARÃES, José Augusto; CANDIDO, Gilberto. The terminological dimension of provenance description in the Multilingual Archival Terminology – ICA: some translation problems [presentation]. In: ASSOCIAZIONE ITALIANA PER LA TERMINOLOGIA (ASS.I.TERM): INTERNATIONAL CONFERENCE «TERMINOLOGY AND KNOWLEDGE ORGANIZATION IN PRESERVING DIGITAL MEMORIES», 26., APRIL 14TH TO 16TH 2016, AT THE UNIVERSITY CLUB OF THE UNIVERSITÀ DELLA CALABRIA. Rende: Associazione Italiana per la Terminologia : International Society for Knowledge Organization : Laboratorio di Documentazione of the Università della Calabria, 2016 Disponível em: http://www.assiterm91.it/http://www.assiterm91.it/wp-content/uploads/2015/10/The-terminological-dimension-of-provenance-description-in-the.pdf Acesso em: 2017-04-04.

UNIVERSIDADE DO PORTO. **DeltCI: Dicionário Eletrónico de Terminologia em Ciência da Informação**. Disponível em: https://paginas.fe.up.pt/~lci/index.php/1668 Acesso em: 2017-04-04.

WENDT, Igor da Silveira. **Extração de contextos definitórios a partir de textos em língua portuguesa**. 2011. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre. Disponível em: http://hdl.handle.net/10923/1647 Acesso em: 2017-04-04.

WÜSTER, Eugen. **Einführung in die allgemeine Terminologielehre und terminologische Lexikographie**. Berlin: Springer, 1979.

WÜSTER, Eugen; FELBER, Helmut; LANG, Friedrich; WERSIG, Gernot (Eds.). **Terminologie als angewandte Sprachwissenschaft**. München; New York; London; Paris: KG Saur Verlag, 1979.

Recebido/Recibido/Received: 2017-04-04
Aceitado/Aceptado/Accepted: 2018-01-03