

CONTENIDOS DIGITALES: CONVERGENCIA, CONECTIVIDAD, MODELOS Y NUEVAS CARACTERÍSTICAS

Ariel Alejandro Rodríguez García
Coordinador



Z666.7
C66

Contenidos digitales : convergencia, conectividad, modelos y nuevas características / Coordinador Ariel Alejandro Rodríguez García. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.

xvi, 292 p. - (Bibliotecología, información y sociedad)

ISBN: 978-607-30-6167-4

1. Metadatos - Modelos. 2. Datos vinculados. 3. Indexación - Aspectos sociales. 4. Recuperación de información. I. Rodríguez García, Ariel Alejandro, coordinador. II. ser.

Diseño de portada: Nube Magenta

Primera edición: 19 de mayo de 2022

D.R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Instituto de Investigaciones Bibliotecológicas y de la Información

Circuito Interior s/n, Torre II de Humanidades,

pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,

Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-6167-4

Esta edición y sus características son propiedad de la Universidad Nacional Autónoma de México. Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titular de los derechos patrimoniales.

Publicación dictaminada.

Impreso y hecho en México.

Contenido

INTRODUCCIÓN	IX
Ariel Alejandro Rodríguez García	

CONVERGENCIA

PUBLICACIONES CIENTÍFICAS DIGITALES Y EL CICLO DE PRODUCCIÓN DE CONOCIMIENTO EN LAS CIENCIAS SOCIALES	19
Jenny Teresita Guerra González	

METADATOS PARA LA PRESERVACIÓN DIGITAL DE LOS PERIÓDICOS OFICIALES MEXICANOS FEDERALES Y ESTATALES EN LÍNEA.	37
Ángel Villalba Roldán	

ANÁLISE DOS METADADOS DAS TESES DE DOUTORAMENTO NUM REPOSITÓRIO ACADÉMICO: ESTUDO DE CASO DE UM REPOSITÓRIO PORTUGUÊS.	53
Ana Lúcia Terra Gonçalo Brites	

LA UNAM Y SUS MUSEOS UNIVERSITARIOS, SU FUNCIÓN ACADÉMICO-CULTURAL	73
Mariana García Ramírez Andrés Ramírez Aguirre Ariel Alejandro Rodríguez García	

CONECTIVIDAD

METADATOS, CIENCIA DE LOS DATOS Y BIBLIOTECAS	91
Juan Voutssás Márquez	

NUEVAS PERSPECTIVAS DE LOS SISTEMAS DE ETIQUETACIÓN SOCIAL DE LOS CONTENIDOS DIGITALES	113
Ariel Alejandro Rodríguez García	

PRESERVACIÓN DIGITAL Y GESTIÓN DE METADATOS DEL PATRIMONIO CULTURAL EN AMÉRICA LATINA	131
María Camila Restrepo Fernández	
Joel Alhuay-Quispe	

MODELOS

LAS IMÁGENES Y LOS METADATOS EN LAS BASES DE DATOS DE ENCUADERNACIONES HISTÓRICAS.	147
Antonio Carpallo Bautista	

LA APERTURA DE INFORMACIÓN GUBERNAMENTAL COMO PRIMER PASO AL GOBIERNO ABIERTO.	167
Alejandro Ramos Chávez	

EL ACCESO A LA INFORMACIÓN DE ZONAS DE RIESGOS POR EVENTOS HIDROMETEOROLÓGICOS: UNA NECESIDAD DE MODELO DE METADATOS.	183
Juan Pablo Moreno Garduño	
Isnardo Reducindo Ruiz	

LOS SISTEMAS PARA LA ORGANIZACIÓN DEL CONOCIMIENTO EN EL TRATAMIENTO TEMÁTICO DE LOS RECURSOS DE INFORMACIÓN CULTURALES.	201
Adriana Suárez Sánchez	

NUEVAS CARACTERÍSTICAS

DESARROLLO DE PROYECTOS CULTURALES Y ARTÍSTICOS. NUEVOS RETOS DIGITALES O HÍBRIDOS.	219
Juan Ayala Méndez	

UNA MIRADA SOBRE LA DISTRIBUCIÓN DIGITAL DE LA MÚSICA: CARACTERÍSTICAS, EVOLUCIÓN Y RETOS DE LA CULTURA VIRTUAL.	235
Marco Brandão	

ANÁLISIS TERMINOLÓGICO DE LOS ESPACIOS CULTURALES UNIVERSITARIOS CON UNA PERSPECTIVA ARQUITECTÓNICA	251
--	-----

Mariana del Carmen Sánchez Rodríguez
Luis Enrique Sánchez Rodríguez
Catalina Naumis Peña

EL CONTENIDO DIGITAL EN LAS BIBLIOTECAS Y SU ORGANIZACIÓN	271
--	-----

Jorge Gómez Briseño
Guadalupe Vanessa Carolina Gutiérrez Hernández

Análise dos metadados das teses de doutoramento num repositório académico: estudo de caso de um repositório português

ANA LÚCIA TERRA

Universidade de Coimbra, Portugal

GONÇALO BRITES

ADAI-LAETA, Dept. Eng. Mecânica, Universidade de Coimbra, Portugal

INTRODUÇÃO

A partir da última década do século XX, mas sobretudo depois do advento do novo milénio, os repositórios institucionais passaram a ocupar um lugar central no armazenamento, difusão e preservação dos trabalhos académicos de pós-graduação e da investigação produzida pelas instituições de ensino superior.

As normas usadas para descrever estes recursos de informação foram adaptadas e adequadas a este novo contexto, levando à aplicação de novos esquemas de metadados, que passaram a desempenhar novas funções além das tradicionais de representar, organizar e proporcionar acesso aos acervos informacionais. Se as bibliotecas usam normas de descrição únicas e gerais, os repositórios digitais podem optar por um esquema único de metadados ou por um perfil de aplicação onde diferentes elementos de um ou de mais esquemas de metadados são selecionados e combinados num esquema composto (Pavão *et al.* 2015) direitos de acesso, compartilhamento, reutilização, redistribuição e políticas, bem como os requisitos técnicos para visualização, acesso ou preservação de objetos digitalizados ou natos digitais. Este trabalho aborda

questões teóricas sobre esquemas de metadados adotados em repositórios, permitindo manter a qualidade, completeza e consistência de seus dados. Apresenta os elementos Dublin Core (DC).

Neste pressuposto, o modo como os metadados usados num repositório são projetados e aplicados vai determinar a operação do próprio repositório, impactando, por exemplo, na satisfação das necessidades dos utilizadores ou no grau de interoperabilidade com outros repositórios.

Uma perspetiva global e sistemática sobre os metadados existentes num repositório específico poderá fornecer uma visão consistente da realidade, dos problemas e desafios que devem ser considerados para melhorar o serviço existente.

Este trabalho é dedicado ao levantamento e análise do esquema de metadados utilizado nos registos das teses de doutoramento depositadas num repositório institucional. Trata-se de uma abordagem preliminar com os seguintes objetivos de pesquisa:

- a) identificação do esquema de metadados usado nos registos das teses de doutoramento depositadas no repositório da Universidade de Coimbra (Portugal) e das suas variações,
- b) analisar e discutir variações e inconsistências verificadas, considerando estudos publicados sobre outros repositórios,
- c) definir orientações metodológicas e de recolha e análise de dados para um estudo alargado dos metadados da totalidade das teses de doutoramento e das dissertações de mestrado do repositório da Universidade de Coimbra.

Depois de uma breve síntese sobre os principais conceitos operatórios da pesquisa (repositório institucional, teses e dissertações digitais e metadados para estas tipologias documentais), é apresentado o repositório institucional da Estudo Geral da Universidade de Coimbra, repositório escolhido para objeto da nossa investigação. Os resultados preliminares desta pesquisa são apresentados e analisados à luz de outros estudos. Nas conclusões, são sistematizadas algumas ideias para ampliar o trabalho.

REPOSITÓRIOS INSTITUCIONAIS

Os repositórios institucionais podem ser considerados arquivos abertos que servem os interesses dos membros da comunidade académica (professores, investigadores, estudantes, etc.) à qual estão adstritos, recolhendo os seus outputs intelectuais, disponibilizando o acesso continuado, e permitindo a preservação e gestão. Na definição já clássica de Lynch (2003, 328), um repositório institucional universitário é um conjunto de serviços que uma universidade oferece aos membros de sua comunidade para a gestão e divulgação de materiais digitais criados pela instituição e pelos membros da sua comunidade. É essencialmente um compromisso organizacional com a gestão desses materiais, incluindo a preservação a longo prazo, quando apropriado, bem como a sua organização e acesso ou distribuição.

Esta perspectiva é reforçada por Gibbons (2004), a qual defende que um repositório institucional apresenta cinco componentes identitárias: conteúdo digital, foco e orientação para a sua comunidade, enquadramento institucional, continuidade e estabilidade além de conteúdo acessível. O repositório institucional é constituído por conteúdos digitais, sob as mais variadas formas (nato-digitais, digitalizados, texto, áudio, vídeo, imagens, etc.). A comunidade de utilizadores determina o que deve ser depositado no repositório, sendo autores dos conteúdos, ou tendo com eles alguma relação, além de frequentemente ter a responsabilidade de fazer os depósitos. Carece de um compromisso institucional, que garanta financiamento e outros recursos, e que incentive o seu uso pela comunidade. A continuidade e a estabilidade implicam que, ao ser depositado um ficheiro no repositório institucional, este continuará acessível no longo prazo, sendo cumpridas regras de preservação. Adicionalmente, o repositório institucional pressupõe que os seus conteúdos estejam disponíveis para um acesso alargado, e não apenas ao autor do conteúdo, já que está alinhado com o princípio da partilha.

TESES E DISSERTAÇÕES DIGITAIS

As teses e dissertações digitais, conhecidas pela sigla inglesa ETD (Electronic Theses and Dissertations), são documentos digitais que apresentam os resultados da investigação de estudantes de pós-graduação. Tradicionalmente, enquadravam-se na designada literatura cinzenta, que abrangia os documentos não disponíveis nos canais tradicionais de distribuição, sendo, por isso, de difícil acesso (Botelho & Oliveira 2015). Eram, então, documentos em papel, com poucos exemplares, depositados e acessíveis em locais muito circunscritos, normalmente a instituição onde tinham sido apresentadas e na biblioteca nacional do país ou em serviço especializado.

Esta situação começou a mudar em 1987, num encontro em Ann Arbor, no Michigan (EUA), onde se reuniram representantes de uma empresa de serviços de disponibilização de teses e dissertações, de várias universidades americanas (Virginia Tech e Universidade do Michigan) e de duas empresas de *software*. A ideia inicial de usar o SGML foi sendo aperfeiçoada, à medida que a tecnologia evoluía, e, em meados da década de 1990, nos EUA, foi estabelecido que as teses e dissertações deveriam ser entregues pelos estudantes em versão digital (Fineman 2003).

Mas foi na última década do século XX, e sobretudo a partir do novo milénio, que começaram a surgir, e depois predominar, as teses e dissertações nado-digitais, pois as instituições de ensino superior passaram a requerer aos seus estudantes que entregassem ficheiros eletrónicos dos seus trabalhos de doutoramento ou mestrado para obtenção do respetivo grau.

Enquanto documentos digitais, estas teses e dissertações passaram a poder fazer uso da tecnologia digital para incluir vários tipos de elementos dinâmicos como som, vídeos, hiperlinks ou outro tipo de objetos digitais ou aplicações. Outro aspeto distintivo das teses e dissertações digitais, e uma das suas vantagens, reside no facto de terem incrementado a visibilidade da pesquisa e dos trabalhos daí resultantes dos estudantes de pós-graduação. De facto, enquanto um destes trabalhos impresso e depositado na

biblioteca da universidade podia ser consultado duas ou três vezes em toda a sua existência, um documento digital depositado num repositório de uma universidade pode ter milhares de visualizações/*downloads* num só ano (Eden *et al.* 2008).

Para recolher, arquivar, gerir e facultar o acesso às teses e dissertações digitais, as instituições de ensino superior passaram a usar os repositórios institucionais, apoiados em sistemas tecnológicos adequados às especificidades destes documentos.

Embora, o processo global de submissão possa parecer simples para quem o executa, há muitas etapas complexas e uma grande variedade de decisões necessárias para torná-lo fluído e simples (Flynn & Ahrberg 2020). Nas suas diversas componentes técnicas e de gestão, destaca-se o desenvolvimento, adaptação e uso de protocolos de metadados (Park & Richard 2011).

METADADOS PARA TESES E DISSERTAÇÕES DIGITAIS

O desenvolvimento e uso de esquemas de metadados, e a qualidade, consistência e interoperabilidade dos metadados produzidos e disponibilizados, determina fortemente a pesquisa e acesso, bem como a preservação a longo prazo das teses e dissertações (Potvin & Thompson 2016). De facto, uma boa descrição do conteúdo e do contexto das teses e dissertações digitais melhora as potencialidades dos registos serem recuperados em processos de pesquisa dos itens depositados no repositório institucional, além de permitirem outros tipos de uso como por exemplo análises cienciométricas. Neste sentido, Schöpfel (2013) entende que os metadados são uma das cinco componentes que podem acrescentar valor a estes trabalhos académicos, além da qualidade intrínseca da própria tese ou dissertação, do formato (preferencialmente pdf), da interoperabilidade e dos serviços disponibilizados pelo repositório onde estão depositados.

Em 2011, Park & Richard apresentaram um estudo sobre o uso do formato ETD-MS para descrever as teses e dissertações digitais em 10 repositórios institucionais canadianos. Dessa pesquisa

concluíram que se verificavam três variações significativas no uso deste esquema de metadados: 1. Uso de um elemento do ETD-MS com a adição de um qualificador local mas em que o significado permanecia igual ao prescrito pelo ETD-MS; 2. Uso de elementos não qualificados do ETD-MS para fins diferentes dos prescritos pelo ETD-MS; e 3. Uso de elementos do ETD-MS aos quais foram adicionados qualificadores não conformes com as práticas do ETD-MS. Concluíram que existiam diferenças significativas por parte das instituições inquiridas na aplicação do esquema de metadados ETD-MS o que suscitava problemas de normalização e de troca/recolha de dados entre os repositórios.

Numa pesquisa mais recente, Steele & SumCrethar (2016) procuraram fazer um levantamento com o objetivo de identificar os metadados usados nos repositórios das bibliotecas dos EUA para descrever e proporcionar o acesso às suas teses e dissertações. Tendo obtido 51 respostas, concluíram que, nas bibliotecas académicas da sua amostra, há uma grande variação no uso dos elementos incluídos nos registos, afirmando que não existe um esquema padrão para descrever as teses e dissertações nos repositórios institucionais. O título, resumo, autor, assunto e a distinção entre tese e dissertação são os elementos recorrentes mas os restantes estão muito dependentes das escolhas de cada instituição, por vezes, mesmo naquelas que partilham o mesmo repositório institucional.

Num inquérito sobre o modo como as instituições descrevem e proporcionam o acesso às teses e dissertações digitais, respondido por 61 bibliotecas norte-americanas, Flynn & Ahrberg (2020) verificaram que o Dublin Core e o MARC representavam os dois formatos de metadados mais usados, havendo inclusão dos documentos no repositório institucional, para o primeiro, e no catálogo da biblioteca, para o segundo.

De facto, o Dublin Core e o ETD-MS apresentam-se como os dois esquemas de metadados mais usados em repositórios institucionais para criar registos de teses e dissertações digitais.

O Dublin Core Simples, ou não qualificado, é composto por 15 elementos de metadados que se enquadram em três grupos conforme o tipo de dados a que dizem respeito: elementos relativos

ao conteúdo do recurso (Title, Subject, Description, Type, Source, Relation & Coverage), elementos relativos à propriedade intelectual (Creator, Publisher, Contributor & Rights) e elementos relativos à instânciação (Date, Format, Identifier & Language). No Dublin Core Qualificado (Qualified Dublin Core), os elementos refinados tornam o significado de um elemento mais restrito ou mais específico, decompondo o significado de um elemento não qualificado. Para o efeito, aos 15 elementos descritivos do Dublin Core são acrescentados termos qualificadores (“Dublin Core™ Metadata Initiative” 2021).

O ETD-MS é um esquema de metadados concebido, em 2003, pela Networked Digital Library of Theses and Dissertations (<http://www.ndltd.org/>), no intuito de melhorar a interoperabilidade dos metadados, e consequentemente a comunicação entre repositórios, especificamente quanto a teses e dissertações. O ETD-MS é um esquema de metadados baseado no Dublin Core e constituído por 23 elementos. Inclui 13 elementos do Dublin Core não qualificado (Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Type, Coverage & Rights) e um novo elemento, Thesis, incluído nos 10 elementos qualificados (Tabela 1).

Tabela 1. Esquema de metadatos do ETD-MS

Elemento	Tipo de uso
dc.title	Obrigatório repetível
dc.title.alternative	Opcional repetível
dc.creator	Obrigatório repetível
dc.subject	Obrigatório repetível
dc.description	Opcional repetível
dc.description.abstract	Opcional repetível
dc.description.note	Opcional repetível
dc.description.realese	Opcional repetível
dc.publisher	Opcional repetível
dc.publisher.country	

Contenidos digitales...

continuación...	
Elemento	Tipo de uso
dc.contributor	Opcional repetível
dc.contributor.role	Opcional
dc.date	Obrigatório
dc.type	Obrigatório repetível
dc.format	Opcional repetível
dc.identifier	Obrigatório repetível
dc.language	Opcional repetível
dc.coverage	Opcional repetível
dc.rights	Opcional repetível
thesis.degree.name	Opcional repetível
thesis.degree.level	Opcional repetível
thesis.degree.discipline	Opcional repetível
thesis.degree.grantor	Opcional repetível

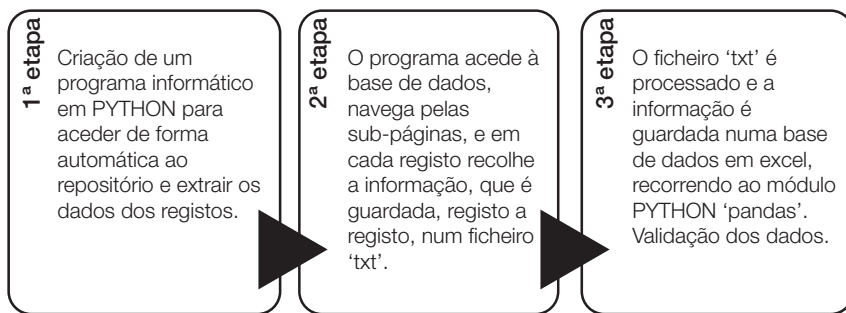
Fuente: adaptado de <http://www.ndltd.org/standards/metadata>.

O ETD-MS apresenta uma tabela de equivalências com os campos do MARC21, permitindo incorporar os registos do repositório nos catálogos das bibliotecas (Hickey, Pavani, and Suleman 2008). A Ken State University (EUA) implementou a integração destes registos no seu catálogo, descrevendo a complexidade do processo de conversão entre o ETD-MS e MARC21 (Eden *et al.* 2008).

METODOLOGIA DA RECOLHA DE DADOS NO ESTUDO GERAL REPOSITÓRIO CIENTÍFICO DA UNIVERSIDADE DE COIMBRA

O Estudo Geral é o repositório digital da produção científica da Universidade de Coimbra, cujo objetivo consiste em divulgar conteúdos digitais de natureza científica de autores ligados à instituição. Foi lançado em 2008, dando cumprimento aos princípios da Declaração de Berlim, que a Universidade de Coimbra tinha subscrito em 2007. A divulgação dos trabalhos de pós-graduação (teses de doutoramento e dissertações de mestrado), sempre que

Figura 1. Procedimentos de recolha de dados



Fuente: elaboração própria.

possível disponibilizando o seu conteúdo em texto integral, constitui um dos objetivos nucleares do Estudo Geral. Assim, o repositório afirma-se como um instrumento para aumentar a visibilidade, acessibilidade e difusão dos resultados da atividade académica da Universidade de Coimbra e da sua investigação científica, de um modo mais geral (Estudo Geral 2018).

Em 24 de novembro de 2020, o Estudo Geral tinha 41.814 itens depositados, nos quais se incluíam 3.975 teses de doutoramento e 18.034 dissertações de mestrado. Portanto, estas duas tipologias documentais representavam metade (52%) dos recursos presentes no repositório. Neste cenário, conhecer o esquema de metadados usado para a incorporação das teses e dissertações no repositório ganha especial significado porque as opções nesta matéria irão determinar a divulgação e a acessibilidade da maioria dos documentos disponibilizados pelo Estudo Geral.

No sentido de delinear uma pesquisa global sobre os metadados das teses e dissertações presentes no repositório da Universidade de Coimbra, foi realizado um estudo piloto e preliminar sobre uma amostra dos registos das teses de doutoramento. De seguida, serão explicitados os procedimentos de recolha de dados no repositório (Figura 1).

Na primeira etapa, foi criado um programa informático para aceder de forma automática à página internet das Teses de Doutoramento

do repositório científico da UC (página principal: <https://estudogeral.uc.pt/browse?type=type&order=ASC&rpp=20&value=doctoralThesis>) e respetivas sub-páginas, retirar a informação de cada um dos registos para uma base de dados, e posteriormente proceder à sua análise. Optou-se por utilizar a linguagem PYTHON devido à sua versatilidade, uma vez que estão disponíveis módulos adicionais que permitem facilmente expandir as capacidades da linguagem às necessidades de todas as áreas científicas.

Na fase seguinte, o programa acede à base de dados, navega pelas sub-páginas, e para cada registo pesquisa pelos campos que queremos recolher, sendo a informação guardada, registo a registo, num ficheiro 'txt'. Isto permitiu confirmar se todos os registos foram extraídos e se todos os campos de cada registo estavam a ser corretamente identificados. Para facilitar o trabalho de identificar as etiquetas 'html' foi utilizado o módulo PYTHON 'BeautifulSoup'.

Na terceira etapa, o ficheiro 'txt' é lido e a informação é guardada numa base de dados, recorrendo ao módulo PYTHON 'pandas'. Como o número de registos não ultrapassa os 2500, optou-se por guardar a base de dados num ficheiro excel, porque tem a vantagem de facilitar a visualização dos dados. Foi necessário proceder à validação dos dados, sendo encontrados registos duplicados, campos sem texto (alguns registos de teses continham informação incompleta), diferentes tipos de separador para as palavras chave, palavras chave em inglês e português misturadas, textos dos resumos em português e inglês juntos no mesmo campo. A validação de dados foi feita, tanto quanto possível de forma automática, utilizando o módulo 'langdetect' para detetar automaticamente a língua, mas foi necessário fazer uma validação manual e correção de alguns dados através de inspeção visual do ficheiro excel.

Seguiu-se o processamento dos dados, tendo sido ponderadas diversas perspetivas para o seu tratamento e análise.

Trabalhou-se com uma amostra de 2496 registos de teses de doutoramento, num universo de 3.975 registos de teses de doutoramento.

APRESENTAÇÃO E ANÁLISE DOS RESULTADOS PRELIMINARES

Os dados recolhidos permitem verificar que, no Estudo Geral, são usados 73 diferentes elementos de descrição das teses de doutoramento (Tabela 2). Neste conjunto, encontram-se 12 dos 15 elementos do Dublin Core, alguns dos quais com qualificadores, quatro elementos do ETD-MS específicos para as teses (thesis.degree.name, thesis.degree.level, thesis.degree.discipline e thesis.degree.grantor) e cinco específicos da Universidade de Coimbra (uc.controloAutoridade, uc.date.periodoembargo, uc.date.periodoEmbargo, uc.degree.grantorUnit, uc.rechabilitacaoestrangeira), entre outros. Estes dados podem ser cotejados com um levantamento realizado por (Steele & Sump-Crethar 2016), em 51 repositórios dos EUA, onde 65% dos inquiridos afirmaram usar o Dublin Core e 26% responderam usar outro esquema de metadados.

Tabela 2. Elenco dos elementos de descrição usados nos registos das teses de doutoramento do Estudo Geral

Designação do elemento		Designação do elemento	
1	dc.contributor.advisor	38	thesis.degree.grantorUnit
2	dc.contributor.author	39	uc.date.periodoembargo
3	dc.date.accessioned	40	dc.relation
4	dc.date.available	41	dc.subject.fos
5	dc.date.issued	42	dc.description.sponsorship
6	dc.identifier.citation	43	dc.date.embargo
7	dc.identifier.uri	44	uc.rechabilitacaoestrangeira
8	dc.description	45	uc.date.periodoEmbargo
9	dc.description.abstract1	46	dc.title.alternative
10	dc.description.abstract2	47	crisitem.author.parentresearchunit
11	dc.description.abstract3	48	dc.date.embargoEndDate
12	dc.language.iso	49	degois.publication.area
13	dc.rights	50	degois.publication.unidadeorganica
14	dc.subject	51	degois.publication.title
15	dc.title	52	dc.rights.uri

Contenidos digitales...

continuación...

Designação do elemento		Designação do elemento	
16	dc.type	53	dc.date.periodoembargo
17	uc.controloAutoridade	54	thesis.degree.disciplineID
18	item.fulltext	55	dc.subject.mesh
19	item.grantfulltext	56	dc.format.mimetype
20	item.languageiso639-1	57	rcaap.embargofct
21	crisitem.advisor.dept	58	dc.publisher
22	crisitem.advisor.parentdept	59	dc.subject.other
23	crisitem.advisor.researchunit	60	thesis.degree.renatesID
24	crisitem.advisor.parentresearchunit	61	bitstream.alternative.ocr
25	crisitem.advisor.orcid	62	dc.description.tableofcontents
26	crisitem.author.dept	63	dc.description.uri
27	crisitem.author.parentdept	64	collection.faculdade
28	crisitem.author.researchunit	65	dc.identifier.isbn
29	crisitem.author.orcid	66	dcterms.subject
30	dc.description.abstract	67	thesis.degree.grantorUnitID
31	dc.date.submitted	68	degois.publication.firstPage
32	degois.publication.location	69	dc.identifier.doi
33	dc.peerreviewed	70	dc.date.created
34	dc.identifier.tid	71	dc.relation.ispartofseries
35	thesis.degree.grantor	72	uc.degree.grantorUnit
36	thesis.degree.level	73	thesis.degree.classification
37	thesis.degree.name		

Fuente: elaboração própria.

Contudo, devemos notar que, se alguns elementos aparecem em todos os registos como, por exemplo, o dc.title, dc.language.iso, dc.contributor.author, dc.date.issued, dc.rights, outros têm ocorrências muito residuais (thesis.degree.classification: 2 vezes, dc.relation: 1 vez, thesis.degree.level: 194 vezes, etc.). Já a identificação do(s) orientador(es) (dc.contributor.advisor), uma informação relevante para efeitos de pesquisa ou de avaliação da produtividade, é inexistente em 35,8% (n. 893) destes registos.

No que respeita aos campos relativos à representação do conteúdo das teses, a saber os resumos e as palavras-chave em português e em inglês, verifica-se a existência de registos sem informação. De facto, se 18,03% (n. 450) dos registos não apresenta palavras-chave em português, este número sobe para 75,20% (n. 1877) no que se refere às palavras-chave em inglês. Esta disparidade nos elementos informativos em português e inglês no mesmo registo, que deveriam ser equivalentes, também acontece no resumo, pois havendo 29,97% (n. 748) de registos que não têm resumo em português este valor sobe para 55,39% (n. 1382) no caso dos resumos em inglês.

Tabela 3. Distribuição do número de palavras nos resumos em português e em inglês

número de palavras	português		inglês	
	%	n.	%	n.
>100	3,9%	68	0,6%	10
100-199	5,1%	89	5,3%	89
200-299	18,9%	330	19,5%	330
300-399	19,1%	334	19,8%	334
400-499	13,2%	230	13,6%	230
500-599	10,2%	178	10,5%	178
600-699	8,6%	150	8,9%	150
700-799	8,8%	154	9,1%	154
800-899	6,1%	106	6,3%	106
900-999	2,5%	44	2,6%	44
1000-1199	1,9%	33	2,0%	33
1200-2000	1,7%	30	1,8%	30
>2000	0,1%	2	0,1%	2
Fuente: elaboração própria.				

No que respeita à extensão dos resumos e ao número de palavras-chave contidas em cada registo também se verificam variações significativas. Assim, nos resumos em português, o registo com a menor extensão, apresenta 16 palavras, e o resumo mais extenso 2801 palavras. Verifica-se que, na amostra, há 1748 regis-

tos com resumos em português, os quais apresentam uma média de 497 palavras por resumo, enquanto a mediana se situa nas 430 palavras por resumo. Analisando com mais detalhe esta distribuição (Tabela 3), constata-se que um pouco mais de metade dos registos apresenta resumos com 200 a 499 palavras e que os registos com menos de 200 palavras ou mais de 899 palavras apresentam valores pouco significativos.

Nos resumos em inglês, verifica-se que a média e a mediana é em ambos os casos de 384 palavras, parecendo haver mais uniformidade quanto à extensão do que nos resumos em português. Constata-se ainda que cerca de 60% dos resumos em inglês apresenta entre 200 a 599 palavras. Os registos com menos de 200 palavras e os que apresentam mais de 899 palavras também constituem percentagens residuais, tal como no português.

Numa análise mais detalhada às palavras-chave, verifica-se igualmente alguma inconsistência nos registos no que respeita ao português e ao inglês (Tabela 4).

Sobressai, desde logo, o facto de haver 2014 registos com palavras-chave em português e apenas 619 registos com palavras-chave em inglês. Esta circunstância pode derivar do facto de algumas teses terem sido elaboradas numa altura em que ainda não era obrigatória a entrega do resumo e das palavras-chave em inglês. Dada esta realidade, foram coletadas 16.671 palavras-chave em português e 5.891 em inglês. Cada registo apresenta em média 3,5 palavras-chave em português, verificando-se a existência de um número muito díspar, variando entre 1 palavra-chave e 57 palavras-chave. Nos registos em inglês, o número médio de palavras-chave é de 5, havendo registos com um mínimo de uma palavra-chave e com um máximo de 35 palavras-chave. Tanto em português (6,4) como em inglês (5,2) o desvio-padrão é bastante acentuado. Estes valores diferem de estudos relativos a outros repositórios, como o repositório da Universidade de São Paulo (Brasil), onde as dissertações de mestrado apresentam em média 4,62 palavras-chave em português e 4,59 palavras-chave em inglês e onde os registos sem palavras-chave ou com apenas uma apresentam percentagens residuais (Terra *et al.* 2020). Outros estudos

que analisaram especificamente palavras-chave em teses de doutoramento também encontraram valores distintos. Na Ohio State University (EUA), Strader (2009) encontrou uma média de 5,9 palavras-chave para cada tese. Já para a University of Illinois em Urbana-Champaign (EUA), Han *et al.* (2016) identificaram uma média de aproximadamente 6 palavras-chave nas teses de doutoramento. Vemos, portanto, que os resultados encontrados na nossa amostra no Estudo Geral precisam de ser verificados na coleção total das teses de doutoramento, considerando a data de submissão do documento, já que, em anos mais antigos, o autor da tese não era obrigado a indicar palavras-chave.

Tabela 4. Distribuição do número de palavras-chave nos registos em português e em inglês

número de palavras-chave	PT		ENG	
	%	n.	%	n.
1	4,1%	84	1,3%	8
2	12,1%	248	2,1%	13
3	10,6%	217	4,8%	30
4	7,7%	157	6,8%	42
5	8,6%	176	8,7%	54
6	7,0%	143	6,1%	38
7	6,2%	126	10,2%	63
8	6,3%	129	8,9%	55
9	4,8%	98	7,4%	46
10	5,3%	108	8,1%	50
11	4,3%	88	7,4%	46
12	4,0%	81	4,4%	27
13	3,0%	62	4,5%	28
14	2,5%	51	3,7%	23
15	1,9%	38	3,9%	24
16	2,1%	42	1,9%	12
17	1,8%	37	2,7%	17

continuación...

número de palabras-chave	PT		ENG	
	%	n.	%	n.
18	1,1%	22	1,5%	9
19	1,4%	28	1,5%	9
20	0,6%	13	0,6%	4
21	0,7%	14	0,6%	4
22	0,4%	9	0,8%	5
23	0,4%	8	0,2%	1
24	0,7%	15	0,2%	1
>24	2,5%	52	1,6%	10
	100,0%	2046	100,0%	619

Fuente: elaboração própria.

Na distribuição do número de palavras-chave por registo, constatam-se novamente diferenças entre o português e o inglês. Nos registos em português, a percentagem mais frequente de palavras-chave é de duas (12,1%) ou três (10,6%), havendo cerca de 60% de registos com entre duas a oito palavras-chave. Já nos registos em inglês, um pouco mais de 60% dos registos apresenta entre quatro a 11 palavras-chave.

CONCLUSÃO

O objetivo deste trabalho centrou-se na apresentação do esquema de metadados usado nos registos das teses de doutoramento disponibilizadas no repositório digital da Universidade de Coimbra e analisar a sua aplicação numa amostra de 2500 registos. Nesse intuito, foi feito um levantamento dos elementos do esquema de metadados usados nos registos das teses de doutoramento. Os resultados evidenciam variações e inconsistências no uso do esquema de metadados do Dublin Core, tendo sido analisados alguns elementos, de um ponto de vista quantitativo e numa abordagem exploratória.

Este estudo preliminar foi realizado no intuito de testar e aperfeiçoar a técnica e os procedimentos de recolha dos metadados no repositório a fim de estabilizar uma metodologia a ser aplicada à totalidade das teses e dissertações depositadas no repositório.

Com os dados obtidos foram realizadas análises que pretendemos ampliar e refinar no corpus geral das teses e dissertações, nomeadamente a frequência de utilização de cada elemento do esquema de metadados aplicado no repositório da Universidade de Coimbra, os padrões de uso de cada elemento e as suas variações, a correlação entre a data do documento e os metadados usados, entre outros.

Verificamos, portanto, que a opção por conceber e implementar um estudo exploratório nos proporcionou bases solidadas para uma pesquisa mais alargada, no que toca à análise dos metadados usados em repositórios institucionais.

REFERÊNCIAS BIBLIOGRÁFICAS

- Botelho, Rafael Guimarães, & Cristina da Cruz de Oliveira. 2015. “Literaturas Branca e Cinzenta: Uma Revisão Conceitual”. *Ciência Da Informação* vol. 44, núm. 3: 501-13. <https://doi.org/10.18225/ci.inf.v44i3.1804>.
- “Dublin Core™ Metadata Initiative”. 2021. <https://www.dublin-core.org/>.
- Eden, Bradford, Sevim Mccutcheon, Michael Kreyche, Margaret Beecher Maurer, & Joshua Nickerson. 2008. “Morphing Metadata: Maximizing Access to Electronic Theses and Dissertations”. *Library Hi Tech* vol. 26, núm. 1: 41-57. <https://doi.org/10.1108/07378830810857799>.
- Estudo Geral. 2018. “Sobre o Repositório”. 2018. <https://estudogeral.uc.pt/sobre.jsp>.
- Fineman, Yale. 2003. “Electronic Theses and Dissertations”. *Portal: Libraries and the Academy* vol. 3, núm. 2: 219-27.

- Flynn, Emily Alinder, & Janet H. Ahrberg. 2020. "Electronic Theses and Dissertations (ETDs) Metadata Policies, Workflows, and Practices: A Survey of the EDT Metadata Lifecycle at United States Academic Institutions". *Journal of Library Metadata* vol. 20, núm. 2-3: 91-110.
- Gibbons, Susan. 2004. "Defining an Institutional Repository". *Library Technology Reports* vol. 40, núm. 4: 6-10. <http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=14019645&site=ehost-live>.
- Han, Myung-Ja K., Patrick Harrington, Andrea Black, & Deren Kudeki. 2016. "Aligning Author-Supplied Keywords for ETDs with Domain-Specific Controlled Vocabularies". In *Classification & Indexing Satellite Conference*.
- Hickey, Thom, Ana Pavani, & Hussein Suleman. 2008. "ETD-MS v1.1: An Interoperability Metadata Standard for Electronic Theses and Dissertations". <http://www.ndltd.org/standards/metadata>.
- Lynch, Clifford. 2003. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age". *Portal: Libraries and the Academy* vol. 3, núm. 2: 328-36.
- Park, Eun G., & Marc Richard. 2011. "Metadata Assessment in E-Theses and Dissertations of Canadian Institutional Repositories". *The Electronic Library* vol. 29, núm. 3: 394-407.
- Pavão, Caterina Groposo, Janise Borges da Costa, Manuela Klano-vicz Ferreira, & Zaida Horowitz. 2015. "Metadados e Repositórios Institucionais: Uma Relação Indissociável Para a Qualidade Da Recuperação e Visibilidade Da Informação". *Ponto de Acesso* vol. 9, núm. 2: 103-16. <https://doi.org/10.9771/1981-6766rpa.v9i3.15163>.
- Potvin, Sarah, & Santi Thompson. 2016. "An Analysis of Evolving Metadata Influences, Standards, and Practices in Electronic Theses and Dissertations". *Library Resources and Technical Services* vol. 60, núm. 2: 99-114. <https://doi.org/10.5860/lrts.60n2.99>.

- Schöpfel, Joachim. 2013. "Adding Value to Electronic Theses and Dissertations in Institutional Repositories". *D-Lib Magazine* 19 (3/4). <http://www.dlib.org/dlib/march13/schopfel/03schopfel.html>.
- Steele, Tom, & Nicole Sump-Crethar. 2016. "Metadata for Electronic Theses and Dissertations: A Survey of Institutional Repositories". In *Journal of Library Metadata*, vol. 16:53-68.
- Strader, C. Rockelle. 2009. "Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations". *Library Resources & Technical Services* vol. 53 núm 4.
- Terra, Ana Lúcia, Carmen Agustín Lacruz, Óscar Bernardes, Mariângela Spotti Lopes Fujita, & Gema Bueno de la Fuente. 2020. "Subject-Access Metadata on ETD Supplied by Authors: A Case Study about Keywords, Titles and Abstracts in a Brazilian Academic Repository". *The Journal of Academic Librarianship*, In press. <https://doi.org/10.1016/j.acalib.2020.102268>.

Contenidos digitales: convergencia, conectividad, modelos y nuevas características. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial Anabel Olivares Chávez; corrección especializada, Valeria Guzmán González; revisión de pruebas Valeria Guzmán González y Carlos Ceballos Sosa; formación editorial, Nube Magenta. Fue impreso en papel cultural de 90 gr. en los talleres de Litográfica Ingramex, Centeno 162-1, Col. Granjas Esmeralda, Alcaldía Iztapalapa, CDMX, C. P. 09810. Se terminó de imprimir en junio de 2022.